

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

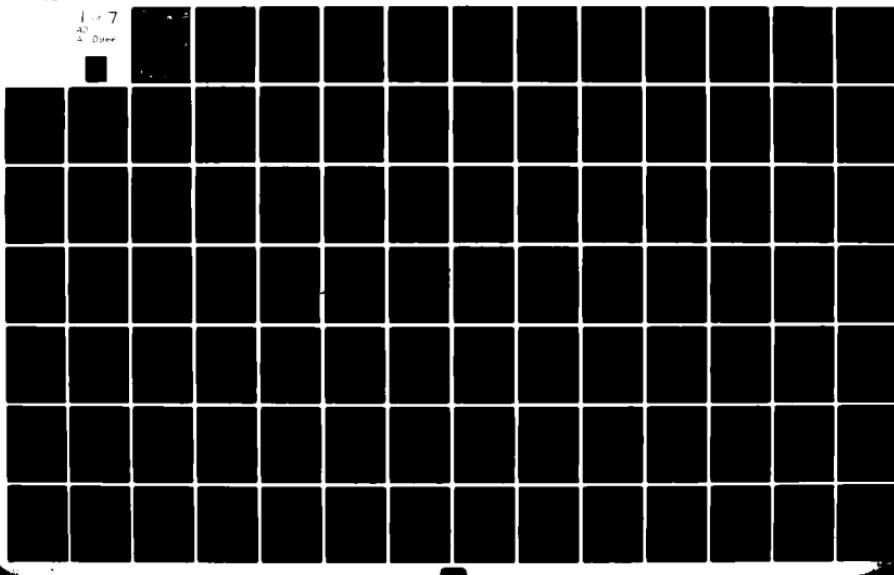
F/G 12/1

UNCLASSIFIED

AFOSR-TR-82-0047

NL

1 of 7  
4D  
2. Date



AFOSR-TR- 82-0047

AFOSR-82-0047

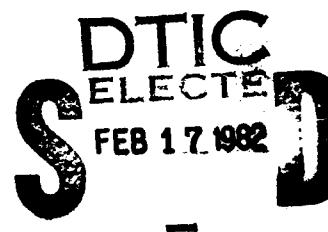
LEVEL II

2

ADA110966

LECTURES ON THE NUMERICAL SOLUTION OF  
PARTIAL DIFFERENTIAL EQUATIONS

PROCEEDINGS OF THE SPECIAL YEAR IN  
NUMERICAL ANALYSIS



held at the

UNIVERSITY OF MARYLAND  
DEPARTMENT OF MATHEMATICS  
COLLEGE PARK, MD 20742

Lecture Notes #20

1981

Edited by  
I. Babuška  
T.-P. Liu  
J. Osborn

DTIC FILE COPY

1982 02 16 170

Approved for public release;  
distribution unlimited.

## UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER <b>AFOSR-TR- 82-0047</b>	2. GOVT ACCESSION NO. <b>ADA110 946</b>	3. RECIPIENT'S CATALOG NUMBER <b>1</b>
4. TITLE (and Subtitle) LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS	5. TYPE OF REPORT & PERIOD COVERED TECHNICAL	
7. AUTHOR(s) I. Babuska, T.-P. Liu, and J. Osborn (editors)	6. PERFORMING ORG. REPORT NUMBER Lecture Notes #20	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Mathematics University of Maryland College Park MD 20742	8. CONTRACT OR GRANT NUMBER(s) AFOSR-80-0251	
11. CONTROLLING OFFICE NAME AND ADDRESS Mathematical & Information Sciences Directorate Air Force Office of Scientific Research Bolling AFB DC 20332	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F; 2304/A3	
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	12. REPORT DATE DEC 1981	
	13. NUMBER OF PAGES 506	
	15. SECURITY CLASS. (of this report) UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The Mathematics Department of the University of Maryland, and the Mathematical and Information Sciences Directorate of the Air Force Office of Scientific research co-sponsored the 1980-81 Special Year in Numerical Analysis. This paper consists of lectures presented by eminent mathematicians from the University and leading visiting mathematicians.		

LECTURES ON THE NUMERICAL SOLUTION OF  
PARTIAL DIFFERENTIAL EQUATIONS

PROCEEDINGS OF THE SPECIAL YEAR IN  
NUMERICAL ANALYSIS

held at the

UNIVERSITY OF MARYLAND  
DEPARTMENT OF MATHEMATICS  
COLLEGE PARK, MD 20742

Lecture Notes #20

1981

Accordance by

N. N. Krasovskii  
V. A. Il'in  
V. A. Kondratenko  
G. I. Marchuk

M. V. Fedorov  
A. G. Sveshnikov

Edited by  
I. Babuška  
T.-P. Liu  
J. Oshorn

AIN FABRICANT  
NOTE  
THE  
APPENDIX  
REFERENCES  
MACINTOSH, INC.  
Chief, Technical Division, Computer Division

TABLE OF CONTENTS:

Preface . . . . .	iii
Special Year Visitors . . . . .	iv
Adapting Courant-Friedrichs-Levy to the 1980's; <i>Garrett Birkhoff</i> . . . . .	1
Parametrization Methods for Approximation of Solutions of Elliptic Boundary Value Problems; <i>J. H. Bramble</i> . . . . .	24
Two Mixed Finite Element Methods for the Simply Supported Plate Problem; <i>J. H. Bramble and Richard S. Falk</i> . . . . .	31
Approximation of Non-linear Problems; <i>F. Brezzi</i> . . . . .	46
Two-dimensional Approximations of Three-dimensional Models in Nonlinear Plate Theory; <i>Philippe G. Ciarlet</i> . . . . .	92
Changing Meshes in Time-dependent Problems; <i>Todd Dupont</i> . . . . .	119
Alternating-direction Galerkin Methods for Parabolic, Hyperbolic and Sobolev Partial Differential Equations; <i>Richard E. Ewing</i> . . . . .	123
Galerkin Methods for Miscible Displacement Problems with Point Sources and Sinks-Unit Mobility Ratio Case; <i>Richard E. Ewing and Mary Fanett Wheeler</i> . . . . ,	151
Tracking of Interfaces in Fluid Flow: Accurate Methods for Piecewise Smooth Problems <i>James Glimm</i> . . . . .	175
Overtaking of Shock Waves in Steady Two-dimensional Supersonic Flows; <i>Ling Hsiao and Tong Zhang</i> . . . . .	204
Homogenization, Convex Analysis, and the Geometry Optimization of Engineering Structures; <i>Robert V. Kohn and Gilbert Strang</i> . . . . .	240
Stability and Error Bounds for a Fractional Step Scheme to Compute Weak Solutions of the Nonlinear Waterhammer Problem; <i>Mitchel Luskin and Blake Temple</i> . . . . .	270
Schauder Estimates for Finite Element Approximations of Second Order Elliptic Boundary Value Problems; <i>Joachim A. Nitsche</i> . . . . .	290
Analysis of Some Contact Problems in Nonlinear Elasticity; <i>J. T. Oden</i> . . . . .	344

Single Step Methods for Linear Differential Equations in Banach Spaces; <i>Vidar Thomée</i>	354
Quasi-optimality of the $H^1$ Projection Into Finite Element Spaces: a Brief Survey with Emphasis on the Maximum Norm (First lecture) <i>Lars B. Wahlbin</i>	390
The Quasi-optimality in the Maximum Norm of the $H^1$ Projection into Piecewise Linear Functions in the Plane: a Complete Proof (Second lecture); <i>Lars B. Wahlbin</i>	398
A Brief Survey of Parabolic Smoothing and How It Affects a Numerical Solution: Finite Differences and Finite Elements (Third lecture); <i>Lars B. Wahlbin</i>	420
Asymptotic Convergence of Boundary Element Methods (First lecture) <i>Wolfgang L. Wendland</i>	435
Integral Equation Methods for Mixed Boundary Value Problems (Second lecture) <i>Wolfgang L. Wendland</i>	478
Defect Correction, Multigrid, and Selected Applications <i>Burton Wendroff</i>	529
Galerkin-Finite Element Solution of Nonlinear Evolution Problems <i>Milos Zlámal</i>	552

## PREFACE

Each year the Department of Mathematics of the University of Maryland sponsors a "Special Year" in some field of mathematics. These special years are designed around a series of lectures by distinguished mathematicians and have the goal of refining the understanding of the frontier of the field, stimulating new research, and enhancing scientific cooperation. During the 1980-81 academic year the Special Year was in numerical analysis. One of the major topics of the Year was the numerical solution of partial differential equations.

Thirty visitors delivered lectures on numerical PDE, touching on nearly all of the important subfields of the area. In addition, many of the participants submitted written versions of their lectures; these papers are contained in this volume. The papers range from extended abstracts of lectures to systematic survey articles to research papers. We have prepared this volume to record the activities of the Special Year and also in the expectation that others will find the papers of interest.

The Organizational Committee would like to thank the Mathematics Department and the Air Force Office of Scientific Research\* for their support, and all of the participants for their stimulating lectures and their informal contribution to the lively scientific climate that prevailed during the Year.

I. Babuška  
T.-P. Liu  
J. Osborn

\*The Special Year was partially supported by AFOSR Grant No. 80-0251.

## PARTICIPANTS

Dr. Garth Baker  
Department of Mathematical Sciences  
State University of New York  
Center at Binghamton  
Binghamton, New York 13901

Professor Garrett Birkhoff  
Department of Mathematics  
Harvard University  
2 Divinity Avenue  
Cambridge, Massachusetts 02138

Professor James Bramble  
Department of Mathematics  
Cornell University  
Ithaca, New York 14853

Professor F. Brezzi  
Laboratorio di Analisi Numerica  
Università di Pavia  
Corso Carlo Alberto 5  
27100 Pavia  
Italy

Professor P. G. Ciarlet  
Analyse Numérique, Tour 55, 5e etage  
Université Pierre et Marie Curie  
4, Place Jussieu  
75230 Paris Cedex 05  
France

Professor Jim Douglas, Jr.  
Department of Mathematics  
University of Chicago  
Chicago, Illinois 60637

Professor Todd Dupont  
Department of Mathematics  
University of Chicago  
Chicago, Illinois 60637

Professor B. Enquist  
Department of Mathematics  
University of California, Los Angeles  
Los Angeles, California 90024

Dr. Richard Ewing  
Mobil Field Research Lab.  
P.O. Box 900  
Dallas, Texas 75221

Professor R. Falk  
Department of Mathematics  
Rutgers University  
New Brunswick, New Jersey 08903

Professor P. Garabedian  
Courant Institute  
New York University  
New York, New York 10012

Professor J. Glimm  
Department of Mathematics  
Rockefeller University  
New York, New York 10021

Professor Amiram Harten  
Department of Mathematical Sciences  
Tel-Aviv University  
Tel-Aviv, Ramat-Aviv  
Israel

Professor Ling Hsiao  
Institute of Mathematics  
Academia Sinica of China  
Pekin, China

Professor P. Lax  
Courant Institute  
New York University  
New York, New York 10012

Professor Mitchell Luskin  
School of Mathematics  
University of Minnesota  
Minneapolis, Minnesota 55455

Professor A. Majda  
Department of Mathematics  
University of California, Berkeley  
Berkeley, California 97720

Professor J. Nitsche  
Inst. für Angewandte Math.  
Hermann-Herder-Str. 10  
D-78 Freiburg  
Germany

Professor J. T. Oden  
TICOM  
W. R. Woolrich Building 304  
University of Texas  
Austin, Texas 78712

Professor J. Oliger  
Department of Computer Science  
Stanford University  
Stanford, California 94305

Professor A. Schatz  
Department of Mathematics  
Cornell University  
Ithaca, New York 14853

Professor Ridgway Scott  
Department of Mathematics  
University of Michigan  
Ann Arbor  
Michigan 48109

Professor G. Strang  
Department of Mathematics  
Massachusetts Institute of  
Technology  
Cambridge, Massachusetts 02139

Professor Roger Teman  
Université de Paris  
Centre D'Orsay  
Mathematique, Batiment 425  
91405 Orsay  
France

Professor Vidar Thomée  
Department of Mathematics  
Chalmers University of Technology  
Fach, S-402 20  
Goteborg  
Sweden

Professor Lars Wahlbin  
Department of Mathematics  
Cornell University  
Ithaca  
New York 14853

Professor W. Wendland  
Fachbereich Mathematik  
Technische Hochschule Darmstadt  
Schlossgartenstrasse 7  
D-6100 Darmstadt  
Germany

Professor B. Wendroff  
Los Alamos Science Lab.  
Group T-7  
Mail Stop 233  
Los Alamos  
New Mexico 87545

Professor Mary Wheeler  
Department of Mathematical  
Sciences  
Rice University  
Houston  
Texas 77001

Professor Milos Zlamal  
Technical University  
Dbrancu Miru 21  
Brno  
Czechoslovakia

# ADAPTING COURANT-FRIEDRICH-S-LEWY TO THE 1980'S

Garrett Birkhoff  
Harvard University

## I. Introduction

In 1928, Courant, Friedrichs and Lewy published a now famous paper [6] on the numerical solution of partial differential equations (DE's). In it, they considered difference approximations to the Laplace, biharmonic, heat, and wave equations. Their stated aim was to treat these by difference methods that were applicable to other partial DE's of elliptic, parabolic, and hyperbolic types, respectively. For simplicity, they used rectangular meshes with constant mesh-length  $h$  in space and (for parabolic and hyperbolic DE's) another constant mesh-length  $\Delta t$  in time.

Their main concern was with proving general existence, uniqueness, and convergence theorems, and not with actually solving specific problems. Their intention was to demonstrate that this area of Analysis (partial DE's) could be arithmetized in principle. They did this so well that their article was probably the most influential paper ever written on the numerical solution of partial DE's. In particular, it strongly influenced von Neumann's later thinking.

Today, computer hardware has increased in efficiency by over a large factor ( $10^6$ ?) that the solution of many partial DE's has been arithmetized in practice. Because of this fact, it seems timely to reconsider the methods proposed in [6], and to compare them with other methods that have been proposed subsequently for solving the same partial DE's.<sup>1</sup>

<sup>1</sup>Analogous commentaries on [6] were written in 1961 by Fox, Milner, and Parter; see [6'].

In comparison with many papers published today on the same subject, [6] seems extremely brief. There are 9½ pages in [6] on the Laplace equation, less than 2 pages on the biharmonic equation, about 3.5 on  $u_{tt} = u_{xx}$ , one page on  $u_{tt} = u_{xx} + u_{yy}$ , and then 4 pages of appendices describing various extensions, of which a little more than one page is devoted to  $u_t = u_{xx}$ .

Numerically, the basic difference approximations to the Laplace and wave equations analyzed in [1] are still in widespread use today, but to achieve accurate approximation with mesh refinement in flat regions, one must modify them near the boundary. The main purpose of this note is to find an effective way of doing this. It will emphasize a central problem pointed out in [1], of crucial importance for elliptic PDE's: that of giving numerically the large sparse system of simultaneous linear algebraic equations to which the discretization used in [6] leads to approximate the Laplacian, and due to Runge (1898), given rise. Neither will it discuss round-off errors, also ignored in [1].

## A. THE LAPLACE EQUATION

2. Dirichlet problems

The problem treated most thoroughly in [6] was the Dirichlet problem. Letting  $M_h$  denote the set of all mesh points  $(x_i, y_j) = (ih, jh)$  lying in the domain  $\Omega$  in which such a problem was posed, they assumed [6', p.221] that a known smooth, but otherwise unspecified function  $g(x, y)$  was interpolated in some boundary strip to the given boundary values on  $\Gamma = \partial\Omega$ . Then Runge's 5-point difference approximation  $\nabla_h^2 U = 0$  to the Laplace equation was solved for the values of  $g$  on the boundary of  $M_h$ .

A better way of approximating Dirichlet-type boundary conditions was developed in the late 1930's and early 1940's by Shortley and Weller [12], R.V. Southwell [13], Leslie Fox [8], and others interested in the practical numerical solution of elliptic problems. As in [6], one first overlays domains in  $\mathbb{R}^2$  with a square mesh, and domains in  $\mathbb{R}^3$  with a cubic mesh, very much in the spirit of [6]. One then supplements the set  $M_h$  of points  $(x_i, y_j) = (ih, jh)$  in  $\Omega$  where the mesh lines intersect with the set  $\Gamma_h$  of boundary nodes  $(x_i, v)$  and  $(x, y_j)$  where a single mesh line intersects the boundary. One then solves  $\nabla_h^2 u = 0$  on  $\Omega_h = M_h \cup \Gamma_h$  for the boundary values in  $\Gamma_h$ . This avoids the problem of interpolating to boundary values,<sup>2</sup> and is generally more accurate.

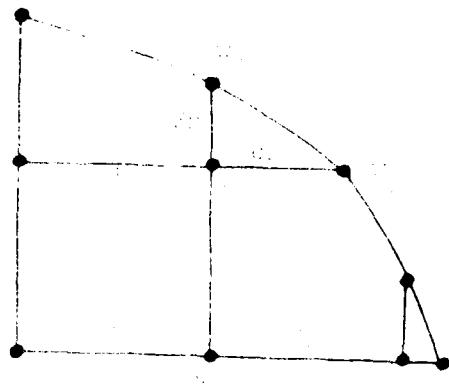
These scientists developed simple and practical difference approximations at mesh points associated with the resulting

---

<sup>2</sup>This problem was considered theoretically in another connection by H. Whitney, Trans. Amer. Math. Soc.

"time dilation" effect. In this paper we will not attempt to determine the exact value of time dilation in the case of the rotating disk, but we will make a rough estimate of likely effects and compare them with the results obtained by other authors. (Instead of the rotating disk we will consider a rotating rectangular plate.

Particularly, we will assume that  $\Omega = \Omega_0$  (const), so that it is not clear which coordinate system will be more appropriate for the second problem, and the choice of the coordinate system will depend on the particular problem. We will also assume that the mass density of the rotating plate is constant, and that the plate is thin enough so that the effects of rotation do not change the mass density distribution. This assumption is not very important for the present problem, since the effects of rotation are small, and the effects of the mass density distribution are also small. However, if the mass density distribution is not constant, then the effects of rotation will be more important, and the effects of the mass density distribution will be more important. This is because the effects of rotation are proportional to the mass density, and the effects of the mass density distribution are proportional to the mass density.



$$G_{\mu\nu} = -\frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W$$

and

$$G_{\mu\nu} = -\frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W$$

$$+ \frac{1}{c^2} \partial_\mu \partial_\nu W = \left\{ \frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W + \frac{1}{c^2} \partial_\mu \partial_\nu W \right\} W$$

The word "rectangular" is used here to mean that the grid is immediately related to a net of rectangular cells, and not to a grid of points. In the former case, the name "rectangular" is not appropriate, since the grid is not rectangular. In any event, no dense interpretation of the grid is implied; the reader should consult the papers of Rutherford and Schrödinger, they will find the explanation below.

Of these two formulas, (2.1) has only  $O(h)$  accuracy, but gives a symmetric matrix; (2.2) is more accurate, but gives an asymmetric matrix.

The symmetry of the matrix given by Shaw's less accurate formula (2.1) is easily explained: it is the formula given by the electrical network analogy, in which each mesh segment is replaced by a conducting wire of the same resistivity per unit length. Note also that when  $\alpha = \beta$ , the two formulas differ only by a factor  $1 + \alpha$ . This may be interpreted physically as corresponding to the area over which the source term  $f(x,y)$  is introducing an inflow of current at the node  $w_0$ .

It is interesting to compare the preceding formulas with the recipe given by Varga in [14, p.186]. Setting  $\alpha = h_i/h$  and  $\beta = k_j/k$  in Varga's (6.37), we get  $(1+\alpha)(1+\beta)$  times Shaw's (2.2). This is encouraging, especially since Varga's (6.37) gives a symmetric matrix in a rectangular domain.

However, in spite of the plausibility of the derivation of Varga's (6.37) in [14, pp.183-5] it seems unlikely that one should use the same weighting for the domain of Fig. 1 as for a rectangle.<sup>4</sup> Moreover, Forsythe-Wasow consider three recipes for boundary conditions in [7, §20.2], and are non-committal as to which is best. It may not even be best to interpolate to boundary values!

---

<sup>4</sup>Well-known monotonicity principles assert, in fact that the weighting factor for  $f$  should increase with the domain. It would be interesting to obtain numerical results for  $-\nabla^2 u = 1$  in the square  $\max(|x|,|y|) \leq 1$  and the octagon satisfying also  $|x| + |y| \leq 3/2$ .

### 3. Normal derivatives

Even less is known about the best way to approximate boundary conditions of the form

$$(3.1) \quad \frac{\partial u}{\partial n} + \alpha(y)u = g(y) \quad \text{on } \Gamma,$$

which was totally ignored in [6], than is known about approximating  $u$ . Thus, whereas it is relatively easy to approximate  $u$  with a truncation error of  $O(h^2)$ , the corresponding error in approximating  $\frac{\partial u}{\partial n}$ , say by

$$(3.2) \quad [\alpha(w_1 - w_0) + \beta(w_2 - w_0)]/\sqrt{\alpha^2 + \beta^2},$$

with the dimensions of Fig. 1, is typically  $O(h)$ .

A brief but incisive summary is given by Forsythe and Wasow in [7, §20.10] of the main ideas and results of Batschelet [?], Shaw [11], Allen [1], and Viswanathan [15].<sup>5</sup> Their summary emphasizes how "complicated" the facts are, and mentions [7, p.204] the possibility of getting improved accuracy by using "reflection" methods. We will next supplement their summary, by discussing some examples.

Example 1. Consider the one-dimensional Poisson DE  $-u''(x) = f(x)$ , with the boundary conditions  $u(0) = u(1) = 0$  and mesh-points  $x_i = ih$ ,  $i = 0, \dots, I$ ,  $h = 1/I$ . Use the Størmer-Numerov approximation.

$$(3.3) \quad u_{i+1} - 2u_i + u_{i-1} = h^2[f_{i-1} + 10f_i + f_{i+1}]/6,$$

---

<sup>5</sup>Batschelet's paper seems the most thorough. The other authors cited take Fox [7] and Southwell [13] as their starting point, and fail to correlate their results with Batschelet's.

a formula whose truncation error is  $h^6 u^{vi}(\xi)/240$  if  $f \in C^4[0,1]$ .<sup>6</sup>

For smooth  $f$  and Dirichlet-type boundary conditions, one can achieve  $O(h^4)$  accuracy with (3.3).

However, with the boundary conditions  $u(0) = 0$  and  $u'(1) = 1$ , the approximation

$$(3.4) \quad u_I = u_{I-1} + h \quad \text{to} \quad u'(1) = 1$$

gives only  $O(h)$  accuracy! If we approximate the boundary condition  $u'(0) = 0$  by

$$(3.5) \quad \begin{aligned} u_I &= u_0 + hu'_0 + h^2 f_0'/2 + h^3 f'_0/6 + O(h^4) \\ &= u_0 + hu'_0 + h^2 (2f_0' + f_1')/6 + O(h^4), \end{aligned}$$

and set  $u_I = 1$ , (3.3) gives  $O(h^2)$  accuracy.

Example 2. Likewise, for the reduced Helmholtz DE,

$$(3.6) \quad u_{xx} = -u_{yy} + \lambda u,$$

the [third] boundary conditions along the line  $y = 0$  can be well approximated on a square mesh by using 9-point formulas in [5] and [10]. If one takes as unknowns the  $u_{-1,j}$  and  $u_{0,j}$ , one gets one equation for each  $j$  from (3.6), and a second equation by collocation from the boundary condition.

$$(3.6') \quad u_x(0,y) + g(y)u(0,y) = h(y).$$

---

<sup>6</sup>See F.H. Hildebrand, Introduction to Numerical Analysis, 2d ed., McGraw-Hill, 1971.

Reflection methods. The preceding method for achieving higher-order accuracy in discretizing boundary conditions is a special application of reflection principles stemming from Fourier (1822), and extended by H.A. Schwarz (ca. 1880) and many others. These are especially applicable to boundary conditions of the form  $u = 0$  or  $\partial u / \partial n = 0$  on straight boundary segments making angles of  $\pi k / 4$  with the x-axis, where  $k$  is an integer. Some simple examples of such applications to the wave equation are presented in Appendix A, "Discretizing Initial and Boundary Conditions."

## B. THE WAVE EQUATION

4. Wave equation: regular mesh

We consider next the semi-discretized wave equation on a square or cubic mesh of side  $h$ :

$$(4.1) \quad u_{tt} = c^2 v_h^2 u,$$

where  $v_h^2$  is the  $(2p+1)$ -point discretized Laplacian in  $p$  space dimensions. We will call a polygonal domain with sides that are all horizontal, vertical, or make a  $45^\circ$  angle with the axes a regular domain when its corners can all be made to fall on mesh-points of such a square or cubic mesh.

The simplest full (central) discretization of (4.1) is

$$(4.2) \quad \delta_{tt} u_j^n = r^2 h^2 \Delta_h^2 u_j^n,$$

where  $r = c\Delta t/h$  is a dimensionless parameter today called the Courant number. The condition for stability is  $r \leq 1/\sqrt{p}$ , and the most accurate stable  $r$  is also the maximum stable  $r$ , with  $r^2 = 1/p$ . This choice reduces (4.2) to the  $(2p+2)$ -point formula

$$(4.3) \quad u_j^{n+1} = \Delta u_j^n - u_j^{n-1},$$

where  $\Delta$  denotes the sum taken over all mesh-points adjacent to  $x_j$ . A 1975 study by Dougalis and the author<sup>†</sup> showed that, in free space, the CFL discretization (4.3) was more efficient than any other second-order discretization, and competitive with later fourth-order schemes.<sup>‡</sup>

---

V.A. Dougalis and G. Birkhoff, pp.231-51 of J.W. Schot and N. Salvesen (eds.), Proc. First International Conference on Numerical Ship, Hydrodynamics, N.S.R.D.C., 1975.

---

<sup>†</sup>L. Collatz, pp.41-61 in J.J. Miller (ed.), Topics in Numerical Analysis, Academic Press, 1973; M. Ciment and S.H. Leventhal, Math. Comp. 29 (1975), pp.985-94.

However, none of the papers referred to above considered in detail how to handle boundary conditions. For boundary conditions of the special form  $u = 0$  on  $\Gamma$  and  $\frac{\partial u}{\partial n} = 0$  on  $\Gamma$ , and more generally for 'mixed' boundary conditions in which some one of these is specified on each edge of a regular domain subdivided by a regular (square or cubic) mesh, we can use a reflection method, stemming from Fourier and applied to the Laplace equation by H.A. Schwarz, to treat boundary conditions without loss of accuracy. Indeed, for  $u = 0$  (the natural physical boundary condition for vibrating membranes, it suffices to set  $u_j^n = 0$  on  $\Gamma$ . For  $\frac{\partial u}{\partial n} = 0$ , a more elaborate procedure is described in the Appendix attached.

### 5. Approximating boundary conditions

In [6], only the pure initial value problem was considered; we next describe a method for approximating behavior near the boundary of a vibrating membrane, where  $u = 0$ . For simplicity, we assume that a convex domain  $\Omega$  in  $\mathbb{R}^P$  with boundary  $\Gamma$  has been overlaid with a uniform ( $p=1$ ), square ( $p=2$ ), or cubic ( $p=3$ ) mesh. This will give rise in general to irregular stars at nodes adjacent to the boundary.

Since the lengths  $\Delta x_i$  of mesh segments adjacent to  $\Gamma$  can be arbitrarily small fractions of  $h$ , the Courant stability criterion  $\Delta t \leq \min(\Delta x_i/c)$  can become a severe limitation near the boundary. But, fortunately, one can circumvent this limitation very easily.

Namely, at the centers of such irregular stars, simply replace the usual hyperbolic difference approximation to  $u_{tt} = c^2 \nabla^2 u$  by the elliptic difference approximation to  $\nabla^2 u = 0$ . In physical language, this amounts to stiffening the membrane artificially at such points, all of which will be adjacent to the boundary. Since  $u = 0$  on  $\Gamma$ , whence  $\nabla^2 u + k^2 u = 0$  implies  $\nabla^2 u = 0$  there, the resulting error should be small except for wave lengths  $\lambda \leq 5h$  (say) very high frequency sound waves. Moreover, it can be reduced further by setting  $u_{tt} = c^2 \theta_j^2 \nabla^2 u$ , where  $\theta_j$  is the minimum ratio of  $\Delta x_i/h$  for a mesh segment issuing from  $x_i$ .

For example, consider the case  $p = 1$ , with domain  $\Omega = [0, \theta h + Jh]$ ,  $0 \leq \theta \leq 1$ . At regular mesh-points

$$(5.1) \quad x_{j+1} = \theta h + jh, \quad j = 1, 2, \dots, J-1,$$

the semi-discretized wave equation reduces to

$$(5.2) \quad u_j''(t) = (c^2/h^2)[u_{j-1} - 2u_j + u_{j+1}].$$

Especially since  $u_0''(0) = u_{tt}(0,t) \equiv 0$  implies  $u_{xx}(0,t) = 0$ , it seems reasonable to approximate

$$u_1(t) = u(\theta h, t) \text{ by } u(h+\theta h)/(1+\theta).$$

For  $\Delta t = h/c$ , this gives

$$(5.3) \quad u_2^{n+1} = \frac{\theta}{1+\theta} u_2^n + u_3^n - u_2^{n-1}$$

and

$$(5.3') \quad u_j^{n+1} = u_{j-1}^n + u_{j+1}^n - u_j^{n-1} \quad \text{for } j > 2.$$

We next estimate the discretization error resulting from the preceding approximation.

Error estimate. One way to estimate the discretization error of (5.3) is to calculate the 'forcing term' required to make the functions

$$(5.4) \quad \phi_k(x,t) = \sin \frac{k\pi x}{J+\theta} \begin{cases} \cos \\ \sin \end{cases} \frac{k\pi t}{J+\theta},$$

which constitute a basis of simply harmonic solutions of  $u_{tt} = u_{xx}$ , become solutions  $U_j^n$  of (5.3) with this term added.

Since the difference and differential equations are time-independent, we can suppose  $t = 0$  without losing generality. Moreover, for the  $\sin$  factor in (5.4), all terms, and hence the forcing term needed to correct for the error, are zero when  $t = 0$ .

There only remains the cos factor, for which

$$(5.5) \quad \delta_{tt} \phi_k = -4 \sin^2 \frac{k\pi h}{2J+2\theta} \sin \frac{k\pi \theta}{J+\theta}.$$

On the other hand, evaluating (5.3), we see that its solution  $U_1$  without a forcing term satisfies

$$(5.5') \quad \begin{aligned} \delta_{tt} U_1 &= \frac{\theta}{1+\theta} \delta_{tt} U_2 \\ &= -\frac{4\theta}{1+\theta} \sin^2 \frac{k\pi h}{2J+2\theta} \sin \frac{k\pi(1+\theta)}{J+\theta}. \end{aligned}$$

The left-hand factors in (5.5) and (5.5') are the same. Expanding the right (spatial) factor of (5.5), we get:

$$(5.6) \quad \sin \frac{k\pi \theta}{J+\theta} = \frac{k\pi \theta}{J+\theta} - \frac{1}{6} \left( \frac{k\pi \theta}{J+\theta} \right)^3 + \dots$$

as compared with

$$(5.6') \quad \frac{\theta}{1+\theta} \sin \frac{k\pi(1+\theta)}{J+\theta} = \frac{k\pi \theta}{J+\theta} \left\{ 1 - \frac{k^2 \pi^2 (1+\theta)^2}{(J+\theta)^2} + \dots \right\}.$$

The forcing term  $f_2^n$  required to make

$$(5.7) \quad U_2^{n+1} = \frac{\theta}{1+\theta} U_2^n + U_3^n - U_2^{n-1} + f_2^n$$

satisfied by  $\phi_k$  is thus  $O(1/J^3)$ ; it is small. This suggests that the local relative order of accuracy of (5.3) at  $x = \theta h$  is  $O(h)$ . Since this is only one of  $J$  mesh points, and the difference equation (5.3') is satisfied exactly elsewhere. The global order of accuracy should be  $O(h^2)$ .

Unfortunately, it seems to be much harder to find a good way to discretize the boundary condition  $\partial u / \partial n = 0$  for a general domain  $\Omega$  with curved boundary  $\Gamma$ . Since this is the boundary

condition that is appropriate for the reflection of sound waves, it would be most desirable to invent a good procedure for discretizing it which would not greatly reduce the maximum stable time step.

### 6. Burgers' and Korteweg de Vries' equations

It is natural to wonder whether prescriptions like those given in §§4-5 have satisfactory analogs for variants of the linear, constant-coefficient wave equation (4.1). For the one-dimensional heat conduction equation  $u_t = u_{xx}$  as well as for (4.1), if  $u(-x,t)$  is a solution then so is  $u(-x,t)$ . It is because of this that solutions satisfying the boundary condition  $u(0,t) \equiv 0$  can be constructed by extending initial conditions anti-symmetrically by the formula

$$(6.3) \quad u(-x,0) = -u(x,0),$$

and  $u_x(0,t) \equiv 0$  can be built into a solution by the following symmetric extension of initial data:

$$(6.3') \quad u(-x,0) = u(x,0).$$

For the Burgers equation (6.1), it is still true that if  $u(x,t)$  is a solution, then so is  $-u(-x,-t)$ . Hence, we can still satisfy the boundary condition  $u(0,t) \equiv 0$  by using the extended initial condition (6.3). However, one cannot 'force' the condition  $u_x(0,t) \equiv 0$  by an analog of 6.3.

For the Korteweg de Vries equation, which was originally proposed as a higher-order nonlinear approximation to 'simple' gravity waves moving in one direction, one cannot satisfy either type of boundary condition by reflection symmetry. This is because, for the transformation  $x \mapsto -x$ ,  $t \mapsto t$ ,  $u \mapsto \lambda u$ , to respect (6.2) for general initial data, we must have  $\lambda = -\lambda^2 = \lambda^2$ . Hence neither  $u(0,t) \equiv 0$  nor  $u_x(0,t) \equiv 0$  can be satisfied by reflecting the initial conditions.

## REFERENCES

- [1] D.N. de G. Allen, Relaxation Methods, McGraw-Hill, 1954.
- [2] E. Batschelet, "Über die numerische Auflösung von Randwertproblemen...", Z. ang. Math. Phys. 3 (1952), 165-93.
- [3] G. Birkhoff and R.E. Lynch, Numerical Solution of Elliptic Problems. SIAM Publications, 1982.
- [4] J. Bramble and B.H. Hubbard, "Approximation of solutions of mixed boundary value problems for Poisson's equation...", J. Assoc. Comp. Mach. 12 (1965), 14-23. (See also SIAM J. Num. Anal. 2 (1965), 1-14).
- [5] L. Collatz, Numerical Treatment of Differential Equations, 3d ed. Springer, 1960.
- [6] R. Courant, K. Friedrichs, and H. Lewy, "Über die partielle D'gleichungen der mathematischen Physik," Math. Annalen 100 (1928), 32-74.
- [6'] Translation of [6] by Phyllis Fox, with comments by Peter Lax, Seymour Parter, and Olof Widlund, IBM J. Res. (1967), 215-47.
- [7] G.E. Forsythe and W. Wasow, Finite Difference Methods for Partial Differential Equations, Wiley, 1960.
- [8] Leslie Fox, "The numerical solution of elliptic partial differential equations...", Phil. Trans. Roy. Soc. A242 (1950), 345-78.
- [9] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial Value Problems. Wiley, 1967.
- [10] L.V. Kantorovich and V.I. Krylov, Approximate Methods of Higher Analysis (translation by Curtis Benster), Interscience, 1958.
- [11] F.S. Shaw, An Introduction to Relaxation Methods, Dover, 1953.
- [12] G.H. Shortley and R. Weller, "The Numerical solution of Laplace's equation," J. Appl. Phys. 9 (1938), 336-48.
- [13] R.V. Southwell, Relaxation Methods in Theoretical Physics, vol. 1, Clarendon Press, 1946.
- [14] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, 1962.

[15] R.V. Viswanathan, "Solution of Poisson's equation...normal gradient specified on curved boundaries," Math. Tables Aids Comp. 11 (1957), 67-78.

## APPENDIX. DISCRETIZING INITIAL AND BOUNDARY CONDITIONS

In the famous paper [1] by Courant, Friedrichs, and Lewy, one of the most important contributions was their 7-point formula for discretizing the two dimensional wave equation

$$(1) \quad u_{ij}^{n+1} = c^2(u_{xx}^{n+1} + u_{yy}^{n+1}).$$

For the optimal Courant number  $\alpha = 1/\sqrt{2}$ , this is

$$(2) \quad u_{ij}^{n+1} = \frac{1}{\pi} \Delta u_{ij}^n - u_{ij}^{n-1}.$$

It is 'hard to beat', because it is explicit, has  $O(h^2)$  accuracy with  $\Delta t = \Delta x/c\sqrt{2}$ , and requires only 4 additions (and subtractions) and one binary shift per time step. However, their formula (2) does not explain how to handle initial or boundary conditions.

Initial conditions. As regards initial conditions, one always has a superposition of two cases:

$$(3a) \quad u_{ij}^n(0) = f_{ij}, \quad \dot{u}_{ij}^n(0) = 0,$$

and

$$(3b) \quad u_{ij}^n(0) = 0, \quad \dot{u}_{ij}^n(0) = g_{ij},$$

respectively. In the first case, we can use the method of reflection: since  $u_{ij}^n(t) = u_{ij}^n(-t)$ , we can replace (2) when  $n = 0$  by

$$(4a) \quad u_{ij}^1 = \frac{1}{\pi} \Delta u_{ij}^0 - u_{ij}^0.$$

In the second case, we know by the reversibility of (1) that  $u_{j\ell}(-t) = -u_{j\ell}(t)$ . Hence  $D^2u(x,y;0) = 0$ , and we can logically replace (2) when  $n = 0$  by

$$(4b) \quad u_{j\ell}^1 = g_{j\ell}\Delta t + O(h^3).$$

Therefore, using the Whittaker or Birkhoff-Lynch [3] to infer  $\dot{u}(x,y;0)$  from the  $g_{j\ell}$ , one can presumably achieve higher-order accuracy in estimating the  $u_{j\ell}^1$  from the data (4b). This would require applying known (exact) Green's functions and their derivatives to the interpolant thus obtained.

Boundary conditions. We will consider here only the case of a polygon with horizontal, vertical, and  $45^\circ$  lines as edges, for Dirichlet-type and/or Neumann-type boundary conditions.

The case of Dirichlet-type boundary conditions,  $u_{j\ell}(t)$  given on  $\Gamma$ , is very easy. The only alteration is that, in Eq. (2), one or more of the terms in  $\Delta u_{j\ell}^n$  is a known quantity (function of time), whenever  $u_{j\ell}^n$  is adjacent to the boundary.

For Neumann-type boundary conditions, one must however use the method of reflection across the boundary. Thus, if  $(x_1, y)$  is an interface (vertical side of the polygon), we must set  $u_{0\ell}^n = u_{2\ell}^n$  on that side. Substituting into (2), this gives after cancellation,

$$(5a) \quad u_{1\ell}^{n+1} = u_{1,\ell+1}^n + u_{\ell-1}^n - u_{1\ell}^{n-1}.$$

Likewise, if  $(x_i, y_{m-i})$  lies on the oblique interface  $x + y = mh$ , then we must set  $u_{i+1,m-i}^n = u_{i,m-i-1}^n$ . Setting  $m = 2$  and  $i = 1$ , this replaces (2) by

$$u_{11}^{n+1} = 2(u_{01}^n + u_{10}^n - u_{11}^n) - u_{11}^{n-1}, \text{ for example, or } \delta_t^2 u_{11}^n = 2(u_{01}^n + u_{10}^n).$$

Rules like the preceding cover all boundary points (where the values of  $u$  must be treated as unknowns for the boundary condition  $\partial u / \partial n = 0$  on  $\Gamma$ ), except corners. Here one must consider six cases: (a)  $90^\circ$  corner formed by horizontal and vertical edges, (b)  $90^\circ$  corner formed by two diagonal edges, (c)  $135^\circ$  corners, (d)  $225^\circ$  corners, (e)  $270^\circ$  corner formed by horizontal and vertical edges, and (f)  $270^\circ$  corner formed by two diagonal edges. Our recommendations for these cases are as follows:

Case 1. A  $90^\circ$  corner between horizontal and vertical edges. Without loss of generality, we can take these edges to be the horizontal and vertical axes. The configuration of Fig. 1a shows how to express the boundary values in terms of interior values.

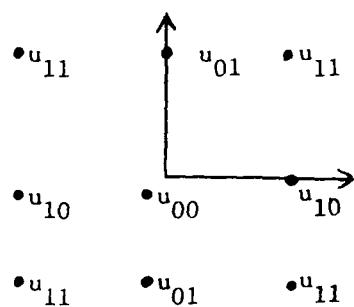


Fig. 1a

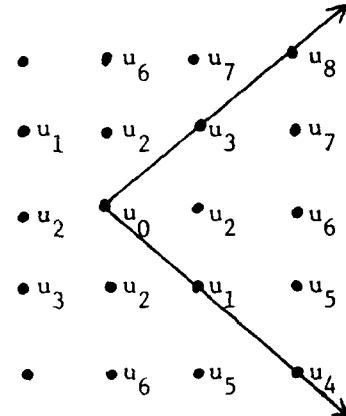


Fig. 1b

Case 2. A  $90^\circ$  corner between two diagonal edges. Without loss of generality, we can assume this is the wedge  $-\pi/4 \leq \theta \leq \pi/4$  depicted in Fig. 1b. Hence we can use reflection to obtain equations for the boundary  $u_k$ , as illustrated in Fig. 1b.

Case 3. Any  $135^\circ$  corner can be transformed by translation, rotation, and reflection into the corner  $y \geq 0$ ,  $x + y \geq 0$  (i.e., into  $\min(y, x+y) \geq 0$ ). The reflections corresponding to the edges  $y = 0$  and  $x + y = 0$  yield the identities  $u_{j,-1} = u_{j,1}$  and  $u_{-j-1,j} = u_{-j,j+1}$ , respectively; see Fig. 2a.

The logic of reflection symmetries involving reentrant corners subtending angles  $\alpha > 180^\circ$  is more subtle. One must in effect imagine a Riemann surface in which the given angle together with its images under reflection in the sides subtends an angle  $180^\circ + \alpha > 360^\circ$ . Of the three cases dual to Cases 1-3, that dual to Case 1 is logically the simplest. By a rotation, we can transform it to the following.

Case 4. Consider a square mesh of side  $h$  that fills the first 3 quadrants, as in Fig. 2b. Reflection in the positive  $x$ -axis, corresponding to the boundary condition  $\partial u / \partial n = \partial u / \partial y = 0$ , suggests setting  $u(h,-h) = u_{11}$ ; reflection in the negative  $v$ -axis suggests that we should set  $u(h,-h) = u_{-1,-1}$ , which appears to be inconsistent. However, the inconsistency between these two formulas is only apparent, and can be resolved by thinking of the origin as a branch point.

in a three-sheeted Riemann surface. In polar coordinates:

$$u(r, -\theta) = u(r, \theta) \quad u(r, 3\pi - \theta)$$

whence

$$u(r, \theta) = u(\theta - 3\pi) = u(\theta - 6\pi).$$

From a computational standpoint, the relevant difference equations are

$$\begin{aligned} u_{10} &= u_{00} + 2u_{11} + u_{20} \\ u_{0,-1} &= u_{00} + 2u_{-1,-1} + u_{0,-2} \\ u_{00} &= u_{10} + u_{-10} + u_{0,-1} + u_{-1,1}. \end{aligned}$$

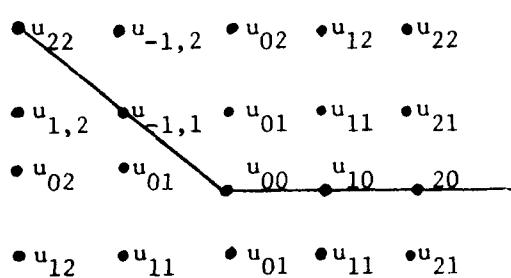


Fig. 2a

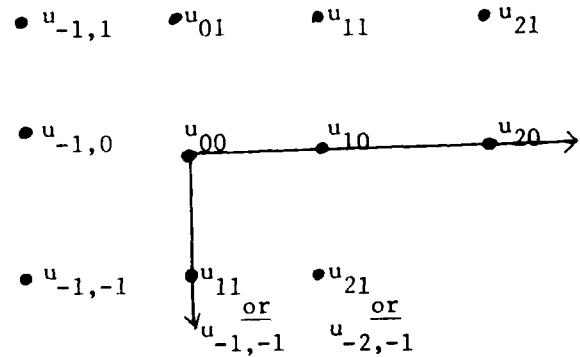


Fig. 2b

Case 5. The complement of case 2 can be treated similarly.

By a translation and a rotation, we can transform it to the domain  $\Omega : -45^\circ \leq \theta \leq 225^\circ$ . It is a good exercise to number the mesh points in this sector near the corner sequentially, and

then write out the equations which express the edge values  $u_{j,-j}$  and  $u_{-j,-j}$  as linear combinations of interior values. The hardest case to treat is the vertex value ( $u_0$  in Fig. 1b); it is not clear that replacing  $u_2$  by the average of the two reflected values will give a suitable answer.

Case 6. The complement of Case 3 leads to a similar difficulty!

Parametrization Methods for Approximation of Solutions  
of Elliptic Boundary Value Problems

by

J. H. Bramble  
Cornell University

The purpose of this talk is to reconsider the Lagrange multiplier method introduced by Babuška [2] and to present some new error estimates as well as a rapidly convergent iteration for the computation of the solution. One of the main points which I wish to make is that the approach given here applies quite well to many other problems. Problems which can be treated by similar methods include interface problems, exterior problems, scattering problems, the Stokes equations and the elasticity equations, the biharmonic problem (with first and second type boundary conditions) and the polyharmonic Dirichlet problem. I will illustrate the results and the approach here by discussing a second order model problem, and the biharmonic Dirichlet problem. A complete discussion of the second order problem may be found in [4].

Let  $\Omega$  be a bounded domain in  $d$ -dimensional space  $\mathbb{R}^d$  with smooth boundary  $\partial\Omega$ . Consider the Dirichlet problem for the Laplacian

$$1) \quad \begin{aligned} Lu &= -\Delta u = f \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega . \end{aligned}$$

For  $\alpha > 0$ , set

$$A(\phi, \psi) = \sum_{j=1}^d \int_{\Omega} \frac{\partial \phi}{\partial x_j} \frac{\partial \psi}{\partial x_j} dx + \alpha \langle \phi, \psi \rangle,$$

where  $\langle \phi, \psi \rangle = \int_{\partial\Omega} \phi \psi ds$ . Define

$$Tf = v,$$

where

$$Lv = f \text{ in } \Omega$$

and

$$\frac{\partial v}{\partial n} + \alpha v = 0 \text{ on } \partial\Omega$$

and

$$G\sigma = \omega,$$

where

$$L\omega = 0 \text{ in } \Omega$$

and

$$\frac{\partial \omega}{\partial n} + \alpha \omega = \sigma \text{ on } \partial\Omega.$$

Here  $\partial/\partial n$  is the outward normal derivative on  $\partial\Omega$ . Now write

$$2) \quad u = Tf + G\sigma$$

where  $u$  is the solution of 1). Note that  $\sigma = \frac{\partial u}{\partial n} + \alpha u$  on  $\partial\Omega$ .

We can, loosely speaking, formulate 1) as follows: Find  $\sigma$  such that

$$G\sigma = g - Tf \text{ on } \partial\Omega.$$

Then 2) gives the solution of 1).

We seek now an approximation of  $u$  in a subspace  $S_h \subset H^1(\Omega)$  which we call  $u_{kh}$ . To define  $u_{kh}$  we first define an approximation  $\sigma_k$  in  $S_k \subset L_2(\partial\Omega)$  as an approximation to  $\sigma$ . This we do, in turn, by approximating  $T$  and  $G$  by projecting onto  $S_h$  relative to  $A(\cdot, \cdot)$ .

(as an inner product on  $H^1(\Omega)$ ). Thus we define (cf. [3])

$$T_h = P_1 T \text{ and } G_h = P_1 G$$

where  $P_1$  is the  $H^1$ -projection given by  $A(P_1 \varphi - \varphi, x) = 0$  for all  $\varphi \in H^1(\Omega)$  and  $x \in S_h$ . Let  $p_0$  be the  $L_2(\Omega)$ -projection onto  $S_k$ . Then we define  $\sigma_k$  by  $p_0 G_h \sigma_k = p_0 (g - T_h f)$  and

$$u_{kh} = G_h \sigma_k + T_h f.$$

Now it is easy to see that

$$A(u_{kh}, \varphi) = (f, \varphi) + (\sigma_k, \varphi)$$

and

$$(u_{kh} - g, x) = 0$$

for all  $\varphi \in S_h$  and  $x \in S_k$  where  $(\cdot, \cdot)$  is the  $L_2(\Omega)$  inner product.

These are essentially the same as the equations given by Babuška [2].

The main stability estimate (proved in [4]) is the following. If  $h \leq \varepsilon k$ , for  $\varepsilon$  sufficiently small and fixed, we have that

$$3) \quad C_0 |\theta|_{-1/2}^2 \leq \langle G_h \theta, \theta \rangle \leq C_1 |\theta|_{-1/2}^2.$$

Here  $C_0$  and  $C_1$  are constants independent of  $\theta$ ,  $h$  and  $k$ . We have tacitly assumed the usual approximation properties for  $S_h$  and  $S_k$  and inverse properties are required only for the spaces  $S_k$  (cf. [3]). Now if  $r$  and  $\dot{r}$  are the parameters indicating the degree of approximation of  $S_h$  and  $S_k$  respectively, then we can state, for example, the following error estimates:

$$|\sigma - \sigma_k|_{-3/2} + \|u - u_{kh}\| \leq C(h^r \|u\|_r + k^{\dot{r}+3/2} |\sigma|_{\dot{r}}).$$

Here the norms are the indicated norms in the appropriate Sobolev spaces and  $\|\cdot\|$  is the  $L_2(\Omega)$  norm. Furthermore if  $|\cdot|$  denotes the  $L_2(\partial\Omega)$ -norm and if  $\dot{S}_k \subset H^{r-3/2}(\partial\Omega)$  then

$$|\frac{\partial u}{\partial n} - (\sigma_k - \alpha g)| \leq C(k^{-r+1/2} h^{2r-2} \|u\|_r + k^r |\sigma|_r).$$

Both these estimates are new and may be found in detail in [4].

We may illustrate the extension of this technique on the biharmonic Dirichlet problem

$$\Delta^2 u = f \text{ in } \Omega$$

4)

$$u = \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega.$$

Set

$$u = T^2 f + TG\sigma.$$

Then

$$\Delta^2 u = f \text{ in } \Omega$$

and

$$\frac{\partial u}{\partial n} + \alpha n = 0 \text{ on } \partial\Omega$$

for any  $\alpha$ . Hence determine  $\sigma$  such that

$$TG\sigma = -T^2 f \text{ on } \partial\Omega.$$

Then  $u = 0$  and thus  $\frac{\partial u}{\partial n} = 0$  so that  $u$  is the solution to 4). The approximation is now clear. Set

$$u_{kh} = T_h^2 f + T_h G_h \sigma_k$$

with

$$p_0 T_h G_h \sigma_k = -p_0 T_h^2 f.$$

The analogous stability and error estimates are

$$5) \quad c_0 |\theta|_{-3/2}^2 \leq \langle T_h G_h \theta, \theta \rangle \leq c_1 |\theta|_{-3/2}^2$$

and

$$|\sigma - \sigma_k|_{-3/2} + \|u - u_{kh}\| \leq C(h^r \|u\|_r + k^{r+5/2} |\sigma|_r).$$

Again assumptions similar to those made previously concerning  $S_h$  and  $\dot{S}_k$  are tacitly being made. This approximation was given by Falk [6] but the estimates here are new.

We finally consider the question of computing  $\sigma_k$ . For  $\dot{S}_k \subset H^1(\partial\Omega)$  define the "discrete surface Laplacian"  $\ell_k : \dot{S}_k \rightarrow \dot{S}_k$  by

$$\langle \ell_k \phi, \chi \rangle = \langle \phi, \chi \rangle_1$$

for all  $\phi, \chi \in \dot{S}_k$ . Here  $\langle \cdot, \cdot \rangle_1$  is an inner product on  $H^1(\partial\Omega)$ . Now  $\ell_k$  is positive definite and symmetric and hence  $\ell_k^S$  is defined in the usual way. Now it can be shown [4] that on  $\dot{S}_k$

$$c_0 |u| \leq |\ell_k^{1/4} u|_{-1/2} \leq c_1 |u|.$$

This together with the stability estimate 3) yields

$$6) \quad c_0 |u|^2 \leq \langle (\ell_k^{1/4} p_0 G_h \ell_k^{1/4}) u, u \rangle \leq c_1 |u|^2$$

for some constants  $c_0$  and  $c_1$ . This means that the matrix induced by  $\ell_k^{1/4} p_0 G_h \ell_k^{1/4}$  has a uniformly bounded condition number and hence in order to solve the system

$$p_0 G_h \sigma_k = p_0 (g - T_h f)$$

we solve instead

$$7) \quad (\ell_k^{1/4} p_0 G_h \ell_k^{1/4}) \theta = \ell_k^{1/4} p_0 (g - T_h f)$$

and then obtain  $\sigma_k$  by

$$\sigma_k = \ell_k^{1/4} \theta.$$

Equation 7) may be solved efficiently by the conjugate gradient method (cf. [1]) because of 6). We have assumed that the operator  $\ell_k^{1/4}$  is easy to compute which may be the case. For example if  $d = 2$  and  $\dot{S}_k$  consists of periodic smoothest splines on a uniform partition of  $\partial\Omega$  then  $\ell_k^{1/2}$  may be obtained by using the fast Fourier transform [5] in  $O(k^{-1} \ln k^{-1})$  operations. When this is not so easy to compute, other stability estimates given in [4] lead also to efficient computational procedures.

In the case of the biharmonic problem we want to solve

$$8) \quad p_0 T_h G_h \sigma_k = -p_0 T_h^2 f.$$

The estimate 5) and the properties of  $\ell_k$  lead to

$$c_0 |\mu|^2 \leq (\ell_k^{3/4} p_0 T_h G_h \ell_k^{3/4})_{\mu, \mu} \leq c_1 |\mu|^2$$

for  $\mu \in \dot{S}_k$ . This leads us to formulate 8) as

$$9) \quad (\ell_k^{3/4} p_0 T_h G_h \ell_k^{3/4})_{\mu} = -\ell_k^{3/4} p_0 T_h^2 f$$

with

$$\ell_k = \ell_k^{3/4} \mu.$$

As in the case of 7), equation 9) may be solved efficiently using the conjugate gradient method.

30

- [1] Axelsson, O., Solution of linear systems of equations: iterative methods, Sparse Matrix Techniques, V.A.Barker (editor), Lecture Notes in Mathematics 572, Springer-Verlag, 1977.
- [2] Babuška, I., The finite element method with Lagrangian multipliers, Numer. Math. 20 (1973), 179-192.
- [3] Bramble, J. H. and Osborn, J. E., Rate of convergence estimates for nonselfadjoint eigenvalue approximations, Math. Comp. 27 (1973), 525-549.
- [4] Bramble, J. H., The Lagrange multiplier method for Dirichlet's problem, Math. Comp. 37 (1981).
- [5] Cooley, J. W. and Tukey, J. W., An algorithm for the calculation of complex Fourier series, Math. Comp. 19 (1965), 297-301.
- [6] Falk, R. S., A Ritz method based on a complementary variational principle, R.A.I.R.O. Analyse numérique 10 (1976), 39-48.

Two Mixed Finite Element Methods  
for the  
Simply Supported Plate Problem

by

James H. Bramble  
Department of Mathematics  
Cornell University  
Ithaca, N.Y. 14853

and

Richard S. Falk  
Department of Mathematics  
Rutgers University  
New Brunswick, N.J. 08903

Talk presented at the University of Maryland on April 15, 1981 by  
Richard Falk.

## 1. Introduction.

In this paper we wish to study two mixed finite element methods for the approximation of a boundary value problem modeling a simply supported plate, i.e. we consider the biharmonic equation

$$(1) \quad \Delta^2 \tilde{u} = f \quad \text{in } \Omega,$$

subject to the boundary conditions

$$(2) \quad \Delta u - \nu (\tilde{u}_{ss} + K \tilde{u}_n) = 0$$

and

$$(3) \quad u = 0 \quad \text{on } \Gamma,$$

where  $\Omega$  is a bounded domain in  $\mathbb{R}^2$  with smooth boundary  $\Gamma$ ,  $f$  is a given function,  $K$  is the curvature of  $\Gamma$ ,  $1-\nu$  is Poisson's ratio, and  $\tilde{u}_s$  and  $\tilde{u}_n$  denote the tangential and exterior normal derivatives of  $u$  respectively along  $\Gamma$ .

In the standard variational formulation of (1)-(3), (2) is a natural boundary condition and so the solution  $\tilde{u}$  may be characterized by

$$(4) \quad \begin{aligned} & \text{Find } \tilde{u} \in H^2(\Omega) \cap H_0^1(\Omega) \text{ such that} \\ & \{ (u, v) - \nu \{ (\tilde{u}_{xx}, v_{yy}) + (\tilde{u}_{yy}, v_{xx}) - 2(\tilde{u}_{xy}, v_{xy}) \} = (f, v) \\ & \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega) \\ & \text{(where } (\cdot, \cdot) \text{ denotes the } L_2(\Omega) \text{ inner product).} \end{aligned}$$

If one bases a finite element method on this variational principle, one is faced with the difficulty of constructing subspaces of  $H^2(\Omega) \cap H_0^1(\Omega)$ . This requires the use of  $C^1$  finite

elements which must vanish on  $\partial\Omega$ .

By using the mixed method technique of introducing new independent variables (e.g.  $\tilde{w} = -\Delta \tilde{u}$ ), we are able to reformulate this problem as a lower order system of equations. This will allow us to define a conforming finite element method using only  $C^0$  finite elements. In addition, we make use of the Lagrange multiplier method to handle the problem of essential boundary conditions.

The two finite element methods we shall consider are based on two different variational formulations of Problem (1)-(3). For simplicity, we shall mainly deal in this paper with the simpler variational formulation, valid for domains with strictly positive curvature (i.e.  $K > 0$ ). The case of general  $K$  will be dealt with briefly at the end of the paper.

Let  $\langle \cdot, \cdot \rangle$  denote the  $L^2(\Gamma)$  inner product and also the pairing between  $H^s(\Gamma)$  and  $H^{-s}(\Gamma)$  and let

$$A_\alpha(u, v) = (\text{grad } u, \text{grad } v) + \alpha \langle u, v \rangle$$

where  $\alpha$  is chosen sufficiently large so that  $2\alpha + K > 0$ . We then consider:

Problem  $(\tilde{P})^*$ : Find  $(\tilde{u}, \tilde{w}, \beta) \in H^1(\Omega) \times H^1(\Omega) \times H^{-1/2}(\Gamma)$  such that

$$(5) \quad A_{\alpha}(\tilde{w}, v) = (f, v) + \langle \beta, v \rangle \quad \text{for all } v \in H^1(\Omega),$$

$$(6) \quad A_{\alpha}(\tilde{u}, z) = (\tilde{w}, z) + \langle \frac{\tilde{w}}{K}, z \rangle \quad \text{for all } z \in H^1(\Omega),$$

and

$$(7) \quad \langle \tilde{u}, g \rangle = 0 \quad \forall g \in H^{-1/2}(\Gamma).$$

To understand the relation between Problem  $(\tilde{P}^*)$  and the biharmonic problem (1)-(3) observe first that equation (5) is the weak form of the boundary value problem

$$\begin{aligned}-\Delta \tilde{w} &= f \quad \text{in } \Omega, \\ \frac{\partial \tilde{w}}{\partial n} + \alpha \tilde{w} &= \sigma \quad \text{on } \Gamma,\end{aligned}$$

and equation (6) is the weak form of the boundary value problem

$$\begin{aligned}-\Delta \tilde{u} &= \tilde{w} \quad \text{in } \Omega, \\ \frac{\partial \tilde{u}}{\partial n} + \alpha \tilde{u} &= -\frac{\partial \tilde{w}}{\partial \tilde{\Gamma}} \quad \text{on } \Gamma.\end{aligned}$$

Equation (7) gives the boundary condition  $\tilde{u} = 0$  on  $\Gamma$ .

Suppose now that for  $u$  a smooth solution of (1)-(3) we set

$$(8) \quad w = -\Delta u$$

and

$$(9) \quad \varepsilon = -\left[ \frac{\partial^2}{\partial n^2} (\Delta u) + \alpha \tilde{u} \right].$$

Then by (1.1),  $-\Delta \tilde{w} = f$  and by (8)-(9),  $\varepsilon = \frac{\partial^2}{\partial n^2} \tilde{w} + \alpha \tilde{w}$  which implies that  $(\tilde{u}, \tilde{w}, \varepsilon)$  satisfies (5). Now from (5),  $\tilde{u}_{ss} = 0$  on  $\Gamma$  so that by (2) and (8)  $\frac{\partial \tilde{u}}{\partial n} + \alpha \tilde{u} = -\frac{\partial \tilde{w}}{\partial \tilde{\Gamma}}$ . Hence (6) is satisfied. Finally (3) implies (7) so that  $(\tilde{u}, \tilde{w}, \varepsilon)$  with  $\tilde{w}$  and  $\sigma$  defined by (8)-(9) is a solution of Problem  $(\tilde{P}^*)$ .

Based on this variational formulation, we now consider the following finite element scheme. Although other choices are possible we shall for simplicity let  $S_h$ ,  $0 < h < 1$ , be the restriction to  $\Omega$  of  $\{v \in C^0(\tilde{\Omega}); v|_t \in P_{r-1}, \forall t \in \tau_h\}$  where  $P_{r-1}$  denotes polynomials of degree  $r-1$  or less in  $x$  and  $y$  and  $\tau_h$  denotes a triangulation of some fixed polygon  $\tilde{\Omega}$  containing  $\Omega$  with

triangles  $t$  of diameter  $\leq h$ . For  $0 < k < 1$  we shall denote by  $\dot{S}_k$  the  $\{\sigma \in C^0(\Gamma) : \sigma|_I \in P_{r-1}, \forall I \in \tau_k\}$  where  $P_{r-1}$  denotes polynomials of degree  $r-1$  or less as a function of arclength along  $\Gamma$  and  $\tau_k$  is a quasiuniform partition of  $\Gamma$  into subintervals  $I$  of arclength  $\leq k$ .

With this choice of subspaces, our finite element scheme is given by:

Problem  $\tilde{P}_h^{k*}$ : Find  $(\tilde{u}_h, \tilde{w}_h, \sigma_k) \in S_h \times S_h \times \dot{S}_k$  such that

$$(10) \quad A_\alpha(\tilde{w}_h, v_h) = (f, v_h) + \langle \sigma_k, v_h \rangle \quad \text{for all } v_h \in S_h$$

$$(11) \quad A_\alpha(\tilde{u}_h, z_h) = (\tilde{w}_h, z_h) - \langle \frac{\tilde{w}_h}{\tau K}, z_h \rangle \quad \text{for all } z_h \in S_h$$

and

$$(12) \quad \langle \tilde{u}_h, \beta_k \rangle = 0 \quad \text{for all } \beta_k \in \dot{S}_k.$$

The motivation for this formulation comes from the following ideas.

Define operators

$$T: H^s(\Omega) \rightarrow H^{s+2}(\Omega)$$

and

$$G: H^s(\Gamma) \rightarrow H^{s+3/2}(\Omega)$$

by  $A_\alpha(Tf, v) = (f, v)$  for all  $v \in C^\infty(\bar{\Omega})$

and  $A_\alpha(G\sigma, v) = \langle \sigma, v \rangle$  for all  $v \in C^\infty(\bar{\Omega})$ ,

i.e.  $Tf$  is the weak solution of the boundary value problem

$$-\Delta(Tf) = f \quad \text{in } \Omega$$

$$\frac{\partial}{\partial n}(Tf) + \alpha(Tf) = 0 \quad \text{on } \Gamma$$

and  $G\sigma$  is the weak solution of the boundary value problem

$$-\Delta(G\sigma) = 0 \quad \text{in } \Omega$$

$$\frac{\partial}{\partial n}(G\sigma) + \alpha(G\sigma) = \sigma \quad \text{on } \Gamma.$$

Using these definitions we see from (5)-(6) that

$$\tilde{w} = Tf + G\sigma$$

and

$$\begin{aligned}\tilde{u} &= T\tilde{w} - G\left(\frac{1}{\tau K} \tilde{w}\right) \\ &= T^2 f + TG\sigma - G\left(\frac{1}{\tau K} Tf\right) - G\left(\frac{1}{\tau K} G\sigma\right).\end{aligned}$$

Let us now define

$$u(\sigma) = TG\sigma - G\left[\frac{1}{\tau K} G\sigma\right].$$

Then  $\tilde{u} = T^2 f - G\left(\frac{1}{\tau K} Tf\right) + u(\sigma)$  so that Problem  $(\tilde{P}^*)$  can be stated in the form:

Problem  $(P^*)$ : Find  $\sigma \in H^{-1/2}(\Gamma)$  such that

$$u(\sigma) = -T^2 f + G\left[\frac{1}{\tau K} Tf\right] \quad \text{on } \Gamma.$$

As we shall now show, the approximation scheme Problem  $\tilde{P}_h^{k*}$  can be viewed as an approximation of the above formulation where we approximate the function  $\sigma$  and the operators  $T$  and  $G$ .

Let us define operators

$$T_h: H^{-1}(\Omega) \rightarrow S_h$$

and

$$G_h: H^{-1/2}(\Gamma) \rightarrow S_h$$

by

$$A_\alpha(T_h f, x) = (f, x) \quad \forall x \in S_h$$

and

$$A_\alpha(G_h \sigma, \chi) = \langle \sigma, \chi \rangle \quad \forall \chi \in S_h.$$

These are just the standard Ritz-Galerkin approximations to  $T$  and  $G$ .

Using the operators  $T_h$  and  $G_h$  we can also rewrite Problem  $\tilde{P}_h^{k*}$  in a form analogous to Problem  $P^*$ . From (10) we have that

$$(15) \quad \tilde{w}_h = T_h f + G_h z_k$$

and from (11) that

$$(14) \quad \tilde{u}_h = T_h \tilde{w}_h - G_h \left[ \frac{1}{\tau K} \tilde{w}_h \right] = T_h^2 f - T_h G_h z_k - G_h \left[ \frac{1}{\tau K} T_h f \right] - G_h \left[ \frac{1}{\tau K} G_h z_k \right].$$

We now define for  $z \in H^{1/2}(\dot{\gamma}, \mathbb{R})$

$$(15) \quad u_h(z) = T_h G_h z - G_h \left[ \frac{1}{\tau K} G_h z \right].$$

Then

$$\tilde{u}_h = T_h^2 f - G_h \left[ \frac{1}{\tau K} T_h f \right] + G_h z_k$$

so that Problem  $\tilde{P}_h^{k*}$  can be restated in the form:

Problem  $\tilde{P}_h^{k*}$ : Find  $z_k \in \dot{S}_K$  such that

$$(16) \quad P_0 G_h z_k = -P_0 T_h^2 f + P_0 G_h \left[ \frac{1}{\tau K} T_h f \right],$$

where  $P_0$  is the  $L_2(\dot{\gamma})$  projection into  $\dot{S}_K$ .

The main idea of this formulation is that the system of linear equations corresponding to this operator equation can be solved in an efficient manner using the preconditioned conjugate gradient method. We now examine how this can be done.

To apply the conjugate gradient method we need to be able to compute  $P_0 u_h(\sigma)$  for any  $\sigma \in \dot{S}_k$ . From the definition of  $u_h(\sigma)$  we see that this involves the solution of two Neumann problems involving the same matrix at each iteration. Hence once an initial LU factorization is found, the calculation of  $u_h(\sigma)$  will involve only two backsolutions. The application of  $P_0$  then requires one additional backsolution per iteration after an initial factorization of the matrix corresponding to  $P_0$ .

For this method to be effective we would like to precondition the iteration so that the spectral condition number and hence the number of iterations required will be independent of the mesh size.

Our choice of preconditioning is based on the following result.

Lemma 1: For  $h \leq \varepsilon k$ , with  $\varepsilon$  sufficiently small, there exist positive constants  $C_1$  and  $C_2$  independent of  $\sigma$ ,  $h$ , and  $k$  such that

$$C_1 |\sigma|_{-1}^2 \leq |\langle P_0 u_h(\sigma), \sigma \rangle| \leq C_2 |\sigma|_{-1}^2 \quad \text{for all } \sigma \in \dot{S}_k.$$

To make use of this result we define a discrete boundary Laplacian

$$\ell_k: \dot{S}_k \rightarrow \dot{S}_k \quad \text{by}$$

$$\langle \ell_k \sigma, \theta \rangle = \langle \sigma, \theta \rangle + \langle \sigma_s, \theta_s \rangle$$

for all  $\theta \in \dot{S}_k$ .

It is then possible to show that

$$C_1 |\sigma|_{-1}^2 \leq |\ell_k^{-1/2} \sigma|_0^2 \leq C_2 |\sigma|_{-1}^2$$

for all  $\sigma \in \dot{S}_k$ , where  $C_1$  and  $C_2$  are constants independent of  $k$ .

Inserting this result in Lemma 1 and setting  $\sigma = \ell_k^{1/2} \theta$  we get

$$c_1 |\theta|_0^2 \leq |\langle P_0 u_h (\ell_k^{1/2} \theta), \ell_k^{1/2} \theta \rangle| \leq c_2 |\theta|_0^2.$$

It then follows from the definition of  $u_h$  that

$$c_1 |\theta|_0^2 \leq |\langle \ell_k^{1/2} P_0 [T_h G_h - G_h (\frac{1}{\tau K}) G_h] \ell_k^{1/2} \theta, \theta \rangle| \leq c_2 |\theta|_0^2.$$

The above inequalities imply that the matrix induced by the (self-adjoint) operator

$$\ell_k^{1/2} P_0 [T_h G_h - G_h (\frac{1}{\tau K}) G_h] \ell_k^{1/2}$$

has a condition number which is bounded independent of  $h$  and  $k$ .

Thus we can obtain a solution  $\theta$  to the equation

$$\begin{aligned} & \ell_k^{1/2} P_0 [T_h G_h - G_h (\frac{1}{\tau K}) G_h] \ell_k^{1/2} \theta \\ &= \ell_k^{1/2} [-P_0 T_h^2 f + P_0 G_h (\frac{1}{\tau K}) T_h f] \end{aligned}$$

to within accuracy  $h^r$  by the conjugate gradient method in  $O(\ell n \frac{1}{h})$  iterations. Returning to the untransformed variables, we now need to compute  $P_0 g$  for  $g \in H^{1/2}(\Gamma)$  and  $[T_h G_h - G_h (\frac{1}{\tau K}) G_h] \sigma$  and  $\ell_k \sigma$  for  $\sigma \in \dot{S}_k$ . As described earlier, all of these require only back substitution at each iteration once some initial factorizations are performed.

We now turn our attention to a brief discussion of error estimates for the approximation scheme just described. The basic variable in our formulation is  $\sigma$  and the key result will be to estimate  $\sigma - \sigma_k$  in appropriate norms. Once this is done, estimates for  $\tilde{u} - \tilde{u}_h$  will follow easily from known estimates for  $(T - T_h)f$  and  $(G - G_h)\sigma$ . To see why, recall that

$$\tilde{u} = T^2 f - G(\frac{1}{\tau K}) T f + T G \sigma - G[\frac{1}{\tau K} G \sigma]$$

and

$$\tilde{u}_h = T_h^2 f - G_h \left( \frac{1}{\tau K} \right) T_h f + T_h G_h \sigma_k^{-1} G_h [\frac{1}{\tau K} G_h \sigma_k].$$

Applying the triangle inequality, we get

$$\begin{aligned} \|\tilde{u} - \tilde{u}_h\|_0 &\leq \|[(T^2 - T_h^2)f]\|_0 \\ &+ \|[G(\frac{1}{\tau K})T - G_h(\frac{1}{\tau K})T_h]f\|_0 + \|TG\sigma - T_h G_h \sigma_k\|_0 \\ &+ \|G[\frac{1}{\tau K} G\sigma] - G_h[\frac{1}{\tau K} G_h \sigma_k]\|_0. \end{aligned}$$

We now show how estimates may be derived for a typical term in the above inequality using standard approximation results and a priori estimates. We write

$$TG\sigma - T_h G_h \sigma_k = (TG - T_h G_h)\sigma + (T_h G_h - TG)(\sigma - \sigma_k) + TG(\sigma - \sigma_k).$$

$$\text{Now } (TG - T_h G_h)\sigma = (T - T_h)G\sigma + (T - T_h)(G_h - G)\sigma + T(G - G_h)\sigma.$$

Let us consider the case  $r = 4$  (piecewise cubics). Then

$$\begin{aligned} \|(T - T_h)G\sigma\|_0 &\leq Ch^4 \|TG\sigma\|_4 \leq Ch^4 \|G\sigma\|_2 \leq Ch^4 |\sigma|_{1/2}, \\ \|(T - T_h)(G_h - G)\sigma\|_0 &\leq Ch^2 \|T(G_h - G)\sigma\|_2 \\ &\leq Ch^2 \|(G_h - G)\sigma\|_0 \leq Ch^4 \|G\sigma\|_2 \leq Ch^4 |\sigma|_{1/2}, \end{aligned}$$

and

$$\|T(G - G_h)\sigma\|_0 \leq \|(G - G_h)\sigma\|_{-2} \leq Ch^4 \|G\sigma\|_2 \leq Ch^4 |\sigma|_{1/2}.$$

A similar argument gives

$$\|(T_h G_h - TG)(\sigma - \sigma_k)\|_0 \leq Ch^4 |\sigma - \sigma_k|_{1/2}.$$

$$\text{Finally, } \|TG(\sigma - \sigma_k)\|_0 \leq C \|G(\sigma - \sigma_k)\|_{-2} \leq C |\sigma - \sigma_k|_{-7/2}.$$

Since the other terms in  $\|\tilde{u} - \tilde{u}_h\|_0$  can be estimated in a similar way, the problem is reduced to estimating  $|\sigma - \sigma_k|$  in various norms.

The main ideas involved in these estimates are the following.

Step 1: Derive an a priori estimate for the continuous problem.

We prove that for all  $s \geq 0$ ,

$$C_1 |\sigma|_{-3/2-s} \leq \|u(\sigma)\|_{1/2-s} \leq C_2 |\sigma|_{-3/2-s}$$

where  $u(\sigma)$  is the solution of the biharmonic problem:

$$\Delta^2 u = 0 \quad \text{in } \Omega$$

$$-\frac{\partial}{\partial n} \Delta u - \alpha \Delta u = \sigma \quad \text{on } \Gamma$$

$$-\Delta u + \tau K[u_n + \alpha u] = 0 \quad \text{on } \Gamma.$$

The relationship of this problem to the original one is that we seek a  $\sigma$  such that  $u(\sigma) = -T^3 f + G(\frac{1}{\tau K} Tf)$  on  $\Gamma$ . For this  $\sigma$ ,  $u(\sigma)$  solves (1)-(3).

It is worth noting that this is not a standard biharmonic problem. From the first boundary condition, it seems that  $\sigma$  should look like three derivatives of  $u$  on  $\Gamma$ . The a priori estimate says it acts like two derivatives. This fact is reflected in the error estimates.

Step 2: Derive similar estimates for the approximate problem.

We prove:

Theorem 1: For  $h \leq \varepsilon k$ , with  $\varepsilon$  sufficiently small, there exist positive constants  $C_1$  and  $C_2$  independent of  $\sigma$ ,  $h$ , and  $k$  such that for all  $0 \leq s \leq \min(r-2, r+\frac{1}{2})$

$$C_1 |\sigma|_{-3/2-s} \leq \|P_0 u_h(\sigma)\|_{1/2-s} \leq C_2 |\sigma|_{-3/2-s} \quad \text{for all } \sigma \in \dot{S}_k.$$

Since this is a continuous dependence theorem for the approximate problem with  $C_1, C_2$  independent of  $h$  and  $k$  we can now get error estimates in the standard way.

Step 3: Let  $\tilde{v}_k \in \dot{S}_k$  be an optimal order approximation to  $v$  in appropriate norms. Then by Theorem 1

$$\begin{aligned} & \|v - \tilde{v}_k\|_{k+5/2-s}^2 = \|P_0[u_h(\tilde{v}_k) - u_h(v)]\|_{1/2-s}^2 \\ & \leq \|P_0[u_h(\tilde{v}_k) - u(v)]\|_{1/2-s}^2 + \|P_0[u(v) - u_h(\tilde{v}_k)]\|_{1/2-s}^2. \end{aligned}$$

Now using the definitions of  $u_h(\tilde{v}_k)$  and  $u(v)$  we get

$$\begin{aligned} & \|P_0[u_h(\tilde{v}_k) - u(v)]\|_{1/2-s}^2 \\ & = \|P_0(\frac{1}{h}f + P_0G_h(\frac{1}{h}T_h f)) \\ & \quad + P_0T^2f - P_0G_h(\frac{1}{h}Tf)\|_{1/2-s}^2. \end{aligned}$$

These terms are easily estimated using known results for  $T-T_h$  and  $G_h$ . Further use of those results and estimates for  $\tilde{v}_k$  allow us to estimate the term  $\|P_0[u(v) - u_h(\tilde{v}_k)]\|_{1/2-s}^2$ .

A special case of our final error estimates gives the following result.

Theorem 2: Suppose  $f \in H^{r+\frac{3}{2}}(\omega)$ ,  $\tilde{v} \in H^{\frac{r}{2}}(\mathbb{T})$  with  $3 \leq r \leq \frac{r+5}{2}$ .

Then for  $h \ll k$  with  $\epsilon$  sufficiently small

$$\|\mu - u_h\|_0 \leq C \cdot h^r (\|f\|_{r+\frac{3}{2}} + \|\tilde{v}\|_{r+\frac{5}{2}}) + k^{\frac{r+5}{2}} |\sigma|_{\frac{r}{2}}.$$

In particular if we use continuous piecewise cubics for  $S_h$  and continuous piecewise linear functions for  $\dot{S}_k$ , then  $r=4$ ,  $\frac{r}{2}=2$  and we obtain the estimate

$$\|\mu - u_h\|_0 \leq C \cdot h^4 (\|f\|_1 + \|\tilde{v}\|_{\frac{5}{2}}) + k^{\frac{9}{2}} |\sigma|_{\frac{1}{2}}.$$

To balance these terms we could choose  $h = k^{9/8}$  so that for  $k$  sufficiently small the condition  $h \ll k$  is automatically satisfied.

We conclude this paper with a brief discussion of a finite

element method valid for arbitrary smooth  $K$ . The method is based on the following variational formulation of the biharmonic problem (1)-(3).

**Problem ( $\tilde{P}$ ):** Find  $(\tilde{u}, \tilde{w}, \cdot, \cdot) \in H^1(\Omega) \times H^1(\Omega) \times H^{3/2}(\Gamma) \times H^{-1/2}(\Gamma)$  such that

$$(17) \quad A_\alpha(\tilde{w}, v) = (f, v) + \tau \langle v - t \lambda_S, v_S \rangle \quad \text{for all } v \in H^1(\Omega),$$

$$(18) \quad A_\alpha(\tilde{u}, z) = (w, z) + \tau \langle z - t \lambda_S, z_S \rangle \quad \text{for all } z \in H^1(\Omega),$$

$$(19) \quad \tau [K(\lambda - \alpha \tilde{u}) + \tilde{u}_S] + \tilde{w} + \mu^\infty = 0 \quad \text{for all } \mu \in H^{3/2}(\Gamma), \text{ and}$$

$$(20) \quad \langle \tilde{u}, \beta \rangle = 0 \quad \text{for all } \beta \in H^{-1/2}(\Gamma).$$

To understand the relation between Problem ( $\tilde{P}$ ) and the biharmonic problem (1)-(3), observe first that equation (17) is the weak form of the boundary value problem

$$\begin{aligned} -\Delta \tilde{w} &= f \quad \text{in } \Omega \\ \frac{\partial \tilde{w}}{\partial n} + \alpha \tilde{w} &= \sigma + t \lambda_S \quad \text{on } \Gamma, \end{aligned}$$

and equation (18) is the weak form of the boundary value problem

$$\begin{aligned} -\Delta \tilde{u} &= \tilde{w} \quad \text{in } \Omega \\ \frac{\partial \tilde{u}}{\partial n} + \alpha \tilde{u} &= \lambda \quad \text{on } \Gamma. \end{aligned}$$

Equations (19) and (20) give the boundary conditions

$$\tau [K(\lambda - \alpha \tilde{u}) + \tilde{u}_S] + \tilde{w} = 0 \quad \text{on } \Gamma$$

and

$$\tilde{u} = 0 \quad \text{on } \Gamma.$$

Suppose now that for  $\tilde{u}$  a smooth solution of (1)-(5) we set

$$(21) \quad \tilde{w} = -\alpha \tilde{u},$$

$$(22) \quad \lambda = \frac{\partial \tilde{u}}{\partial n} + \alpha \tilde{u}$$

and

$$(23) \quad \sigma = -[\frac{\partial}{\partial n} \Delta \tilde{u} + \alpha \Delta \tilde{u} + \tau (\tilde{u}_{nss} + \tilde{w}_{ss})].$$

Then from (1),  $-\beta \tilde{w} = f$  and by (21)-(23)

$$\sigma = \frac{\partial}{\partial n} \tilde{w} + \alpha \tilde{w} - \tau \lambda_{ss}$$

which implies that  $(\tilde{u}, \tilde{w}, \lambda, \sigma)$  satisfies (17). Now from (21) and (22), it easily follows that  $(\tilde{u}, \tilde{w}, \lambda, \sigma)$  satisfies (18). Using (2), (21), and (22) we get that  $\tilde{w} + \tau \{u_{ss} + K(-\alpha \tilde{u})\} = 0$  on  $\Gamma$  and so (19) is satisfied. Finally, (5) implies (20) so that  $(\tilde{u}, \tilde{w}, \lambda, \sigma)$ , with  $\tilde{w}, \lambda, \sigma$  defined by (21)-(23) is a solution of Problem  $\tilde{P}$ .

The approximation proceeds as before except now the basic variables are  $\lambda_k$  and  $\sigma_k$ . We then seek  $\lambda_k, \sigma_k \in \dot{S}_k \times \dot{S}_k$  such that (19)-(20) hold for all  $u, v \in \dot{S}_k \times \dot{S}_k$  where  $\tilde{w}$  is replaced by

$$\tilde{w}_h = T_h f + G_h \{ \sigma_k + \tau (\lambda_k)_{ss} \}$$

and  $\tilde{u}$  is replaced by

$$\tilde{u}_h = T_h \tilde{w}_h + G_h \lambda_k.$$

Once again we get a linear system for  $\lambda_k, \sigma_k$  which can be efficiently solved by the conjugate gradient method after we determine the correct preconditioning. To compute the action of the relevant matrix on a vector we need only be able to apply the operators  $T_h$ ,  $G_h$  and  $P_0$ . As before this is quite easy once some initial factorizations are determined.

References

- [1] Axelsson, O., Solution of linear systems of equations: iterative methods, Sparse Matrix Techniques, V. A. Barker (editor), Lecture Notes in Mathematics 572, Springer-Verlag, 1977.
- [2] Babuška, I. and Aziz, A. K., Survey lectures on the mathematical foundations of the finite element method, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz (Editor), Academic Press, New York, 1972.
- [3] Babuška, I., The finite element method with Lagrangian multipliers, Numer. Math. 20, (1973), 179-192.
- [4] Bramble, J. H. and Osborn, J. E., Rate of convergence estimates for nonseladjoint eigenvalue approximations, Math. Comp. 27 (1973), 525-549.
- [5] Bramble, J. H. and Scott, L. R., Simultaneous approximation in scales of Banach spaces, Math. Comp. 32 (1978), 947-954.
- [6] Bramble, J. H., The Lagrange multiplier method for Dirichlet's problem, preprint.
- [7] Ciarlet, P. G. and Raviart, P. A., A mixed finite element method for the biharmonic equation, Symposium on Mathematical Aspects of Finite Elements in Partial Differential Equations, C. DeBoor, Ed., Academic Press, New York, 1974, pp. 125-143.
- [8] Ciarlet, P. G. and Glowinski, R., Dual iterative techniques for solving a finite element approximation of the biharmonic equation, Comput. Methods Appl. Mech. Engrg., 5 (1975), 277-295.
- [9] Falk, R. S., Approximation of the biharmonic equation by a mixed finite element method, SIAM J. Numer. Anal., 15 (1978), 556-567.
- [10] Glowinski, R. and Pironneau, O., Numerical Methods for the first biharmonic equation and for the two-dimensional Stokes problem, SIAM Review, 21 (1979), pp. 167-212.
- [11] Lions, J. L. and Magenes, E., Problèmes Aux Limites non Homogènes et Applications, Vol. 1, Dunod, Paris, 1968.
- [12] Schechter, M., On  $L_p$  estimates and regularity II, Math. Scand. 13 (1965), pp. 47-69.
- [13] Weinstock, R., Calculus of Variations, McGraw-Hill, New York, 1952.

## APPROXIMATIONS OF NON-LINEAR PROBLEMS

F. Breda

Institute of Mathematics Applications  
Università di Trieste  
Institute of Applied Mathematics  
University of Trieste, Italy.

### 2. INTRODUCTION

We summarize here some recent results obtained in [5], [6], [7], [8] on the finite dimensional approximations of "mildly" nonlinear stationary problems depending on one or more parameters, with special attention to the cases of normal limit points and simple bifurcation points. An outline of the paper is the following. In Chapter 1 we present the general form of the continuous and approximate problems with some general considerations on the hypotheses and some examples. In Chapter 2 we present first two basic abstract results (Section 2.1) in a form which is slightly different from the original one, and we give a sketch of the proofs; then (Section 2.2) we apply these results to the case of the branched off nonsingular solutions. In Chapter 3 we tackle the case of simple singular points and we show how the Liapunov-Schmidt procedure can be applied to reduce both the continuous and the approximate problems to the case of mappings  $\mathbb{E}^2 \rightarrow \mathbb{R}$ . In Chapter 4 we deal with the reduced problems  $\mathbb{E}^2 \rightarrow \mathbb{R}$  in the cases of normal limit points (Section 4.1) and simple bifurcation points (Section 4.2). Finally

in chapter 5 we analyze the case of multiple parameters with special attention to the "perturbation parameters" and to the reproduction, by means of the approximate problem, of the whole bifurcation diagram. Still more has to be done in this direction and we present here examples, ideas and problems rather than a general theory.

We give here only some basic ideas on the proofs; for the details we refer to [5], [6], [7], [4]. Related results can be found in [12], [13], [14], [19], [8], [17], [15] and in the references therein contained.

## C H A P T E R I

**1.1.** We shall consider, in the sequel, nonlinear problems of the following type

$$(1.1) \quad F(u, \lambda) - z = u + T'(u, \lambda) - z,$$

where  $G$  is a  $C^2$ -mapping from  $V \times \mathbb{R}$  into  $W$  and  $T$  is a linear compact operator from  $U$  into  $W$ ; we assume that  $V$  and  $W$  are Banach spaces and that the  $n$ -th derivatives of  $G$  are uniformly bounded, i.e.,  $|G^{(n)}| \leq C$ . As a consequence of (1.1), it is natural to look for solutions in  $U \times \mathbb{R}$  instead of  $V \times \mathbb{R}$ , with  $z$ -coordinates equal to the  $u$ -coordinates, namely  $z = u$ .

Assume now that some mappings  $\{T_h\}_{h>0}$  are given in operator  $T_{h_0} \in L(W, U)$  and assume that

$$(1.2) \quad \lim_{h \rightarrow 0} \frac{\|T_h - T_{h_0}\|}{h} L(W, U) = 0.$$

The "discrete problem" will be defined as follows:

$$(1.3) \quad F_h(u, \lambda) - z = u + T_{h_0}'(u, \lambda) - z = 0$$

and our main object will be to compare the set of the solutions of (1.1) in a neighbourhood of a given point  $(u_0, \lambda_0)$  with the set of the solution of (1.3) in the same neighbourhood.

**1.2.** We shall now make some comments on the nature of the problems and give some examples. We remark first that the

basic assumption is the compactness of the operator  $T$ . In the applications, problem (1.1) will rather be written in the form

$$(1.4) \quad \begin{cases} Au + G(u, \lambda) = 0 \\ \text{+ boundary conditions} \end{cases}$$

where  $A$  will be, say, a linear elliptic operator from a functional space  $V$  into its dual space  $V'$ ; we may then assume that the nonlinear mapping  $G(u, \lambda)$  maps  $V \times \mathbb{R}$  into a subspace  $W$  of  $V'$  and that for any  $f \in V'$  the problem:

$$(1.5) \quad \begin{cases} Au = f \\ \text{+ boundary conditions} \end{cases}$$

has a unique solution  $u = Tf$ . We may also assume, without a "serious" loss of generality, that the boundary conditions are homogeneous, so that  $T$  is a linear operator. In this framework, one could roughly say that the assumption " $T$  is compact from  $V$  into  $V'$ " amounts in a mildness requirement on  $G(u, \lambda)$ : in some sense, the application of  $G$  makes  $A^{-1}$  less regularity, but the successive application of  $T$  or  $A^{-1}$  will re-generate, in the end, even more regularity, so that  $Tf$  will be a "regularizing operator." Assume now that we have a method (a program, a code ...) which can be used to solve (1.5) in some approximate way; then we could consider the approximate solution of (1.4). Usually the approximate solutions of problems of type (1.4) are expressed in

the form

$$(1.6) \quad \|T_h f - Tf\|_V \leq ch^k \|Tf\|_V$$

where  $V$  is a suitable functional space of "regular" functions and the exponent  $k$  depends on the "degree" of the approximation and/or on the regularity of  $V$ . In the abstract theory that follows for nonlinear problems, we shall obtain error estimates of the type

$$(1.7) \quad \|u_h - u\|_V \leq c \| (T_h - T) G(u, \lambda) \|_V$$

that should be considered as optimal in the following sense: if one has an estimate of the type (1.6) for the discrete solution of the linear problem (1.5), then (1.7) will provide

$$(1.8) \quad \|u_h - u\|_V \leq ch^k \|TG(u, \lambda)\|_V = ch^k \|u\|_V$$

which means that the (asymptotic) error in the nonlinear problem is as good as the one we have on the linear problem, for the given "method"  $T_h$ .

A particular case which is of great interest in the applications is the following one, that we shall call the "pure Galerkin case." Assume that we are given a bilinear continuous elliptic form  $a(u, v)$  on  $V \times V$  and assume that  $T$  is defined through  $a(u, v)$  by means of

$$(1.9) \quad \left\{ \begin{array}{l} T : f \in V' \rightarrow Tf \in V, \text{ solution of} \\ a(Tf, v) = \langle f, v \rangle \quad \forall v \in V. \end{array} \right.$$

Assume finally that we are given a family  $\{V_h\}$  of closed subspaces of  $V$ , such that

$$(1.10) \quad \forall v \in V \quad \liminf_{h \rightarrow 0} \inf_{v_h \in V_h} \|v - v_h\|_V = 0;$$

$T_h$  can now be defined by

$$(1.11) \quad \left\{ \begin{array}{l} T_h : f \in V' \rightarrow T_h f \in V_h \text{ solution of} \\ a(T_h f, v) = \langle f, v \rangle \quad \forall v \in V_h. \end{array} \right.$$

If the inclusion  $W \subseteq V'$  is compact,  $T_h$  will satisfy the assumption (1.2) and the abstract theory will be applicable. We recall that, in that case, the estimate (1.6) can be written

$$(1.12) \quad \|T_h f - Tf\|_V \leq c \inf_{v_h \in V_h} \|Tf - v_h\|_V,$$

cfr. e.g. [2].

We shall spend a few words now in order to show that, in fact, the pure Galerkin case is not the only interesting case in which the theory can be applied: we shall restrict ourselves, for the sake of simplicity, to a particular example, but we hope that much more general cases may be easily guessed once this one is understood. Consider in a convex polygon

$\Omega \subset \mathbb{R}^2$  the problem

$$(1.13) \quad \begin{cases} -\Delta u + \lambda g(u, u_x, u_y) = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where  $g$  defines a smooth mapping from  $L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega)$  into  $L^2(\Omega)$  (or, if you wish, into  $H^s(\Omega)$ ,  $s > -1$ ). Assume that you have a "code" which solves the linear problem

$$(1.14) \quad \begin{cases} -\Delta w = f & \text{in } \Omega \\ w = 0 & \text{on } \partial\Omega \end{cases}$$

by "mixed" finite elements; this means that for any  $f \in H^{-1}(\Omega)$  and for any given mesh size  $h$  your code will provide  $T_h f = (w_h, p_h)$  where  $w_h$  is an approximation of  $w$  and  $p_h = (p_h^1, p_h^2)$  is an approximation of  $p = \underline{\text{grad}} w$ . Assume in addition that you have error estimates of the type (see e.g. [2])

$$(1.15) \quad \|w - w_h\|_{L^2(\Omega)} + \|p - p_h\|_{(L^2(\Omega))^2} \leq ch \|w\|_{H^2(\Omega)};$$

then you can study the mixed approximation of problem (1.13) with the following setting

$$V = (L^2(\Omega))^3 \equiv \{v = (\varphi, \tau), \varphi \in L^2(\Omega), \tau \in (L^2(\Omega))^2\},$$

$$W = L^2(\Omega),$$

$$G(v, \lambda) = \lambda g(\varphi, \tau_1, \tau_2),$$

and with  $T$  defined as the mapping that to each  $f \in L^2(\Omega)$  associates  $Tf = (\varphi, \tau_1, \tau_2) \in V$  with  $-\Delta\varphi = f$  ( $\varphi \in H_0^1(\Omega)$ ) and  $\tau = \underline{\text{grad}} \varphi$ . The estimate (1.15) yields now

$$(1.16) \quad \|Tf - T_h f\|_V \leq c|h| \|f\|_{L^2(\Omega)}.$$

Hence we may say that an abstract error of the form (1.7) is still optimal for our mixed approximation of the nonlinear problem (1.13).

## C H A P T E R 2

2.1. We shall now give a theorem that will be used systematically in the sequel; the theorem is a minor modification of a result proved in [5], [6]. However, since the original proof of [5], [6] is rather technical and since the two statements do not coincide exactly, we shall also give a sketch of the proof. The following lemma will be "used" in the proof and will also be useful in the sequel.

Lemma 1. Let  $B_1, B_2$  be Banach spaces and let  $F \in C^1(B_1; B_2)$ ; let moreover  $x_0$  be an element of  $B_1$  such that  $F(x_0) = 0$  and  $DF(x_0)$  is an isomorphism from  $B_1$  onto  $B_2$ . If  $\{F_n\}$  is a sequence of mappings which converges to  $F$  uniformly in  $C^1(B_1; B_2)$  then there exist an integer  $\bar{n}$ , a neighbourhood  $U$  of  $x_0$  in  $B_1$  and a constant  $c$  such that for any  $n \geq \bar{n}$  there exists a unique  $x_n \in U$  such that  $F_n(x_n) = 0$  and we have

$$(2.1) \quad \|x_n - x_0\|_{B_1} \leq c \|F_n(x_0)\|_{B_2} = c \|F_n(x_0) - F(x_0)\|_{B_2}.$$

Proof (sketch). Since  $DF_n \rightarrow DF$  uniformly and since  $DF$  is nonsingular at  $x = x_0$  we have that, for  $n$  big enough,  $(DF_n)^{-1}$  is uniformly bounded in a suitable neighbourhood of  $x_0$  independent of  $n$ . Hence  $F_n$  has an inverse function, say  $G_n$ , in a sphere  $S_\rho(F_n(x_0))$  in  $B_2$  with radius  $\rho$  independent of  $n$ . For  $n$  big enough  $\|F_n(x_0)\|_{B_2} < \rho$  and

hence  $x_n = G_n(0)$  exists and is unique in a neighbourhood of  $x_0$ . Let now  $G$  be the inverse mapping of  $F$ ; we have

$$\begin{aligned}
 \|x_n - x_0\|_{B_1} &= \|G_n F_n G_n(0) - G_n F_n G(0)\|_{B_1} \\
 (2.2) \qquad \qquad \qquad &\leq \sup \|DG_n\| \cdot \|F_n G_n(0) - F_n G(0)\|_{B_2} \\
 &= \sup \|DG_n\| \cdot \|0 - F_n(x_0)\|_{B_2}
 \end{aligned}$$

which completes the proof since  $DG_n = (DF_n)^{-1}$  is uniformly bounded.

We are now able to present the main result of this section; we recall first that if  $\phi$  is a  $C^r$  mapping ( $r \geq 1$ ) from  $X \times Y$  into  $Z$ , where  $X, Y, Z$  are Banach spaces and  $(x_0, y_0)$  is a point in  $X \times Y$  such that

i)  $\phi(x_0, y_0) = 0$

ii)  $D_Y \phi(x_0, y_0)$  is an isomorphism from  $Y$  onto  $Z$

then the classical implicit function theorem ensures the existence of a unique mapping  $g(x) \in C^r(X; Y)$ , defined on a neighbourhood  $N$  of  $x_0$  in  $X$  such that  $g(x_0) = y_0$  and  $\phi(x, g(x)) = 0$  in  $N$ .

Theorem. Let  $X, Y, Z$  be Banach spaces and let  $\phi \in C^r(X \times Y; Z)$  for  $r \geq 1$  with  $D^r \phi$  uniformly continuous on bounded subsets; let  $(x_0, y_0)$  be a point in  $X \times Y$  such that conditions i) and ii) are satisfied and let  $g(x)$  be

the implicit function defined by  $\phi$  in the neighbourhood  $N$  of  $x_0$ . Assume that we are given a sequence  $\{\varphi_n\}$  of  $C^r$  mappings from  $X \times Y$  into  $Z$  and assume that  $\varphi_n$  converges to  $\phi$  uniformly in  $C^r$ . Then there exist: a neighbourhood  $U(x_0)$  in  $X$ , a neighbourhood  $U(y_0)$  in  $Y$ , and integer  $\bar{n}$  and a constant  $c$  such that the following properties hold. For any integer  $n \geq \bar{n}$  there exists a unique mapping  $g_n \in C^r(X; Y)$  defined on  $U(x_0)$  with values in  $U(y_0)$  such that

$$(2.3) \quad \varphi_n(x, g_n(x)) = \phi \text{ in } U(x_0).$$

Moreover  $g_n$  converges to  $g$  uniformly in  $C^r$  and we have for any  $m$  integer with  $0 \leq m < r - 1$ :

$$(2.4) \quad \|D^m(g_n(x) - g(x))\|_{L_m(X, Y)} \leq c \sum_{\ell=0}^m \|D^\ell \varphi_n(x, g(x))\|_{L_\ell(X, Z)}$$

uniformly in  $U(x_0)$ .

Proof (sketch). Consider as usual the auxiliary functions  $F(x, y) = (x, \phi(x, y))$  and  $F_n(x, y) = (x, \varphi_n(x, y))$  which are  $C^r$  mappings from  $X \times Y$  into  $X \times Z$ . Proceeding as in the proof of Lemma 1 we have that the inverse functions  $G(x, z) = (x, \psi(x, z))$  and  $G_n(x, z) = (x, \psi_n(x, z))$  exist in a neighbourhood of  $(0, 0)$  and  $(0, \varphi_n(x_0, y_0))$  (resp.) with fixed radius; obviously  $g(x) = \psi(x, 0)$  and setting  $r_n(x) = \psi_n(x, 0)$  (allowed for  $n$  big enough) we get the implicit function for

$\Phi_n$  which satisfies (2.3). Proceeding as in (2.2) we obtain (2.4) for  $m = 0$ . Then (2.4) has to be proved by induction; we sketch the case  $m = 1$ : remark first that by a suitable choice of  $U(x_0)$  and  $U(y_0)$  we may assume that the first derivatives of  $\Phi$  and  $\Phi_n$  are Lipschitz continuous (now  $r \geq 2$ ). Hence we remark that  $D_y \Phi_n(x, g_n(x)) Dg_n(x) + D_x \Phi_n(x, g_n(x)) \equiv 0$  and  $D_y \Phi(x, g(x)) Dg(x) + D_x \Phi(x, g(x)) \equiv 0$  so that:

$$(2.5) \quad \begin{aligned} & D_y \Phi_n(x, g_n(x))(Dg(x) - Dg_n(x)) \\ &= (D_y \Phi_n(x, g_n(x)) - D_y \Phi(x, g(x))) Dg(x) \\ &+ D_x \Phi_n(x, g_n(x)) - D_x \Phi(x, g(x)). \end{aligned}$$

Since  $(D_y \Phi_n(x, g_n(x)))^{-1}$  is uniformly bounded and  $D_y \Phi_n$ ,  $D_x \Phi_n$  are Lipschitz continuous we have

$$(2.6) \quad \begin{aligned} & \|Dg(x) - Dg_n(x)\|_{L(X, Y)} \\ &\leq C \{ \|g(x) - g_n(x)\|_Y + \|D_y \Phi_n(x, g(x)) \cdot Dg(x) + D_x \Phi_n(x, g(x))\|_{L(X, Z)} \} \\ &= C \{ \|g(x) - g_n(x)\|_Y + \|D \Phi(x, g(x))\|_{L(X, Z)} \} \end{aligned}$$

and (2.4) for  $m = 1$  follows from (2.4) for  $m = 0$  which was already proven. If  $r = 1$  formula (2.5) together with the uniform continuity of  $D_x \Phi_n$  and  $D_y \Phi_n$  shows that  $Dg_n(x)$  converges uniformly to  $Dg(x)$  but does not provide bounds of the form (2.4).

Remark. In the statements of Lemma 1 and Theorem 1 we obviously need only that the functions  $F$ ,  $F_h$  and  $r$ ,  $r_h$  (respectively) are defined, continuous, etc. in a neighbourhood of the point  $x_0$  (resp.  $x_0, \lambda_0$ ).

**2.2.** Using Theorem 1 one easily proves the following result on branches of regular solutions.

Theorem 2. Assume that  $F(u, \lambda)$  is a  $C^r$  mapping ( $r \geq 1$ ) from  $V \times \mathbb{R}$  into  $V$ , with  $D^r F$  uniformly continuous and let  $(u_0, \lambda_0)$  be a point where  $F(u_0, \lambda_0) = 0$  and  $D_u F(u_0, \lambda_0)$  is an isomorphism from  $V$  onto  $V$ ; let finally  $u(\lambda)$  be the implicit function  $U(\lambda_0) \rightarrow V$  such that  $u(\lambda_0) = u_0$  and  $F(u(\lambda), \lambda) = 0$  in  $U(\lambda_0)$ . Let  $F_h(u, \lambda), 0 < h \leq \bar{h}$ , be a family of  $C^r$  mappings from  $V \times \mathbb{R}$  into  $V$  which converges to  $F(u, \lambda)$  uniformly in  $C^r$ ; then there exist a neighbourhood  $U(\lambda_0)$ , a neighbourhood  $U(u_0)$ , an  $h_0 > 0$  and a constant  $c$  such that for each  $h \leq h_0$  there exists a unique mapping  $u_h(\lambda)$  from  $U(\lambda_0)$  into  $U(u_0)$  such that  $F_h(u_h(\lambda), \lambda) = 0$  in  $U(\lambda_0)$ . Moreover  $u_h(\lambda)$  converges uniformly to  $u(\lambda)$  in  $C^r(U(\lambda_0), V)$  and we have for any integer  $m$ ,  $0 \leq m \leq r - 1$

$$(2.7) \quad \left\| \frac{d^m}{d\lambda^m} (u_h(\lambda) - u(\lambda)) \right\|_V \leq c \sum_{\ell=0}^m \left\| \frac{d^\ell}{d\lambda^\ell} F_h(u(\lambda), \lambda) \right\|_V$$

uniformly in  $U(\lambda_0)$ .

Remark. When  $F(u, \lambda)$  and  $F_h(u, \lambda)$  have the form (1.1) and (1.3) respectively, (2.7) takes the form

$$(2.8) \quad \left\| \frac{d^m}{d\lambda^m} (u_h(\lambda) - u(\lambda)) \right\|_V \leq c \sum_{\ell=0}^m \left\| \frac{d^\ell}{d\lambda^\ell} (T_h - T) G(u(\lambda), \lambda) \right\|_V$$

that we already discussed in Section 1.2.

### C H A P T E R 3

3.1. Let now consider points  $(u_0, \lambda_0)$  where  $D_{\lambda}P(u_0, \lambda_0)$  is not an isomorphism. Assuming that  $P(u, \lambda)$  has the form (1.1) we see that  $D_{\lambda}P$  has the form  $I + k$  with  $I =$  identity and  $k =$  compact linear operator; hence  $D_{\lambda}P$  is a Fredholm operator of index zero. From now on we shall assume that  $(u_0, \lambda_0)$  is an elliptic point of (1.1) in which

$$(3.1) \quad \begin{cases} L \in D_{\lambda}P(u_0, \lambda_0) \text{ and there is an eigenvalue of} \\ \text{algebraic multiplicity one,} \end{cases}$$

The classical elliptic operator theory will now ensure the existence of  $\varphi_0 \in V$  and  $\varphi_0^* \in V^*$  ( $=$  dual space of  $V$ ) such that:

$$(3.2) \quad L\varphi_0 = 0, \quad \langle L^*\varphi_0^*, v \rangle = 0, \quad \|\varphi_0\|_V = 1, \quad \langle \varphi_0, \varphi_0^* \rangle = 1;$$

$$(3.3) \quad V = \{\varphi_0\} \dot{+} V_2 \text{ where:}$$

$$(3.4) \quad V_2 = \{\varphi_0\}^\perp = \{v | v \in V, \langle v, \varphi_0^* \rangle = 0\} = E(L),$$

$$(3.5) \quad L|_{V_2} \text{ is an isomorphism of } V_2 \text{ onto } V_2.$$

The decomposition (3.3) will be the starting point for the classical Mairanov-Schmidt decomposition that we shall describe in this section. We introduce first the operator  $Q \in L(V, V_2)$  defined by

$$(3.6) \quad Qv = v - \langle v, \varphi_0^* \rangle \varphi_0$$

and we remark that  $F(u, \lambda) = 0$  in  $V$  if and only if

$$(3.7) \quad QF(u, \lambda) = 0$$

$$(3.8) \quad \langle F(u, \lambda), \varphi_0^* \rangle = 0.$$

The decomposition (3.3) will now be used to write the solution  $(u, \lambda)$  in the form

$$(3.9) \quad u = u_0 + \alpha \varphi_0 + v, \quad \alpha \in \mathbb{R}, \quad v \in V_2$$

$$(3.10) \quad \lambda = \lambda_0 + \xi \quad \xi \in \mathbb{R}.$$

The basic idea of the L-S procedure is to use equation (3.7) to eliminate  $v$  in the expression (3.9), as an implicit function of  $\xi$  and  $\alpha$ . More precisely, let us consider the auxiliary mapping

$$(3.11) \quad F(\xi, \alpha, v) = QF(u_0 + \alpha \varphi_0 + v, \lambda_0 + \xi)$$

which is clearly a  $C^r$  mapping from  $\mathbb{R} \times \mathbb{R} \times V_2$  into  $V_2$ .

Obviously  $F(0, 0, 0) = 0$ ; it is easy to check that

$D_v F(0, 0, 0) = L$ , which is an isomorphism from  $V_2$  onto  $V_2$

(3.5); therefore (3.11) defines uniquely  $v = v(\xi, \alpha)$  with  $v(0, 0) = 0$  as an implicit function. Plugging  $v(\xi, \alpha)$  into (3.9) we have that (3.7) is identically satisfied, so that we have to deal with (3.8) only: setting

$$(3.12) \quad f(\xi, \alpha) = \langle F(u_0 + \alpha \varphi_0 + v(\xi, \alpha), \lambda_0 + \xi), \varphi_0^* \rangle$$

we easily check that  $u = u_0 + \alpha \varphi_0 + v(\xi, \alpha)$ ,  $\lambda = \lambda_0 + \xi$  is a solution of (1.1) iff  $(\xi, \alpha)$  satisfies

$$(3.13) \quad f(\xi, \alpha) = 0 \quad (f : \mathbb{R}^2 \rightarrow \mathbb{R}).$$

**3.2.** Assume now that, as in Section 1.1, we are given a family  $F_h(u, \lambda)$  of  $C^r$  mappings which converges uniformly to  $F(u, \lambda)$  in  $C^r$ ; obviously  $F_h(u, \lambda) = 0$  iff  $\langle F_h(u, \lambda), p_0^* \rangle = 0$  and  $\langle F_h(u, \lambda), \varphi_0^* \rangle = 0$ . Using again the decomposition (3.9), (3.11) we can consider the auxiliary mappings

$$(3.14) \quad F_h(\xi, \alpha, v) = QF_h(u_0 + \alpha \varphi_0 + v, \lambda_0 + \xi);$$

since  $F_h$  converges to  $F$  uniformly in  $C^r$  we may apply Theorem 1 which ensures, for  $h \leq h_0$ , the existence of a unique  $v_h$  mapping a neighbourhood of  $(0, 0)$  in  $\mathbb{R}^2$  into a neighbourhood of  $0$  in  $V_2$  such that

$$(3.15) \quad F_h(\xi, \alpha, v_h(\xi, \alpha)) = 0 \quad \text{near } (0, 0).$$

It is now easy to see that  $u = u_0 + \alpha \varphi_0 + v_h(\xi, \alpha)$ ,  $\lambda = \lambda_0 + \xi$  is a solution of (1.3) iff

$$(3.16) \quad f_h(\xi, \alpha) = \langle F_h(u_0 + \alpha \varphi_0 + v_h(\xi, \alpha), \lambda_0 + \xi), \varphi_0^* \rangle = 0.$$

It is clear that  $v_h(\xi, \alpha)$  converges uniformly to  $v(\xi, \alpha)$  in

$C^r$  and hence  $f_h(\xi, \alpha)$  converges uniformly to  $f(\xi, \alpha)$  in  $C^r$ ; a more careful use of the estimates (2.4) leads to the following theorem (cfr. [6], [7] for the detailed proof).

Theorem 3. We have for all  $m$  with  $0 \leq m \leq r - 1$

$$(3.17) \|D^m(v_h(\xi, \alpha) - v(\xi, \alpha))\|_{L_m(\mathbb{R}^2, V)} \leq c \sum_{\ell=0}^m \|D^\ell F_h(\xi, \alpha)\|_{L_\ell(\mathbb{R}^2, V)}$$

$$(3.18) \|D^m(f_h(\xi, \alpha) - f(\xi, \alpha))\|_{L_m(\mathbb{R}^2, V)} \leq c \sum_{\ell=0}^m \|D^\ell F_h(\xi, \alpha)\|_{L_\ell(\mathbb{R}^2, V)}$$

where  $F_h(\xi, \alpha) = F_h(u_0 + \alpha \varphi_0 + v(\xi, \alpha), \lambda_0 + \xi)$ ; moreover if  $\alpha(t), \xi(t)$ ,  $|t| \leq t_0$  is a  $C^r$  curve in a neighbourhood of  $(0, 0)$  and  $\alpha_h(t), \xi_h(t)$ ,  $|t| \leq t_0$  converges to  $\alpha(t), \xi(t)$  uniformly in  $C^r$  we have for all  $0 \leq m \leq r - 1$

$$(3.20) \begin{aligned} & \left\| \frac{d^m}{dt^m} (v_h(\xi_h(t), \alpha_h(t)) - v(\xi(t), \alpha(t))) \right\|_V \\ & \leq c \sum_{\ell=0}^m \left\{ \left| \frac{d}{dt^\ell} (\xi_h(t) - \alpha(t)) \right| + \left| \frac{d^\ell}{dt^\ell} (\alpha_h(t) - \alpha(t)) \right| \right. \\ & \quad \left. + \left\| \frac{d^\ell}{dt^\ell} F_h(\xi(t), \alpha(t)) \right\|_V \right\} = D(h, m, t) \end{aligned}$$

and

$$(3.20) \left\| \frac{d^m}{dt^m} (f_h(\xi_h(t), \alpha_h(t)) - f(\xi(t), \alpha(t))) \right\| \leq D(h, m, t)$$

where  $D(h,m,t)$  is defined in (3.19).

In summary we may say that in a neighbourhood of a simple singular point (that is, a solution  $(u_0, \lambda_0)$  which satisfies (3.1)) the Liapunov-Schmidt procedure and the "uniform convergence" Theorem 1 allow the reduction of both (1.1) and (1.3) to two dimensional problems

$$(3.21) \quad f(\xi, \alpha) = 0; \quad f_h(\xi, \alpha) = 0$$

with  $f_h \rightarrow f$  in  $C^r$  and with estimates of  $|f_h - f|$  of optimal type in terms of  $|F_h - F|$ . From now on we shall essentially concentrate on the two-dimensional problems (3.21) as if they were our original problems. Obviously the various hypotheses that we shall make on  $f(\xi, \alpha)$  can be "translated" into corresponding hypotheses on  $F(u, \lambda)$ , as has been done in [6], [7]; similarly the error bounds obtained in terms of  $f$  and  $f_h$  should be expressed in terms of  $F$  and  $F_h$  by means of (3.9), (3.10) and Theorem 3 (see again [6], [7] for all the details). We finally remark that if  $(u_0, \lambda_0)$  is a simple singular point we have  $f(0, 0) = \frac{\partial f}{\partial \alpha}(0, 0) = 0$ .

## C H A P T E R 4

4.1. We assume in this section that we are given a  $C^r$  mapping  $f(\xi, \alpha) : \mathbb{R}^2 \rightarrow \mathbb{R}$  with  $f(0,0) = \frac{\partial f}{\partial \alpha}(0,0) = 0$  and a family  $f_h(\xi, \alpha)$  of mappings which converges uniformly in  $C^r$  to  $f(\xi, \alpha)$ , as  $h$  tends to zero, in a neighbourhood of the origin. We shall assume, first, that

$$(4.1) \quad \begin{cases} \text{The origin is a normal limit point for } f(\xi, \alpha), \\ \text{that is } \frac{\partial f}{\partial \xi}(0,0) \neq 0. \end{cases}$$

Hence, the implicit function theorem will ensure the existence of a unique  $C^r$  function  $\xi = \xi(\alpha)$  from  $\mathbb{R}$  to  $\mathbb{R}$  such that  $\xi(0) = 0$  and  $f(\xi(\alpha), \alpha) \equiv 0$  near  $\alpha = 0$ . Theorem 1, in turn, will ensure the existence of a unique  $C^r$  mapping  $\xi_h(\alpha)$  in a neighbourhood of the origin such that  $f_h(\xi_h(\alpha), \alpha) \equiv 0$ ; moreover, for  $0 \leq m \leq r - 1$

$$(4.2) \quad \left| \frac{d^m}{d\alpha^m} (\xi_h(\alpha) - \xi(\alpha)) \right| \leq C \sum_{\ell=0}^m \left| \frac{d^\ell}{d\alpha^\ell} f_h(\xi(\alpha), \alpha) \right|.$$

Since  $\frac{\partial f}{\partial \alpha}(0,0) = 0$  it turns out that necessarily  $\frac{d\xi}{d\alpha}(0) = 0$ . We assume now that the origin is a "nondegenerated turning point," that is

$$(4.3) \quad \frac{d^2 \xi}{d\alpha^2}(0) \neq 0.$$

In that case, for  $r \geq 2$ , we have from Lemma 1 that there exists a unique  $\alpha_h^0$  near 0 such that

$$(4.4) \quad \frac{d\xi_h}{d\alpha} (\alpha_h^0) = 0;$$

moreover

$$(4.5) \quad \begin{aligned} |\alpha_h^0| &= |\alpha_h^0 - 0| \leq c \left| \frac{d\xi_h}{d\alpha} (0) \right| \\ &\leq c \left\{ |f_h(0,0)| + \left| \frac{\partial f_h}{\partial \alpha} (0,0) \right| \right\} = D_1(h). \end{aligned}$$

Hence setting  $\xi_h^0 = \xi_h(\alpha_h^0)$  we have

$$(4.6) \quad \begin{aligned} |\xi_h^0| &\leq |\xi_h(0)| + \left| \frac{d\xi_h}{d\alpha} (0) \right| |\alpha_h^0| + o(|\alpha_h^0|^2) \\ &\leq |f_h(0,0)| + o((D_1(h))^2). \end{aligned}$$

We have proven the following theorem.

Theorem 4. If  $(0,0)$  is a normal limit point of  $f(\xi,\alpha)$  then there exist a neighbourhood  $U$  of  $(0,0)$  and an  $h_0 > 0$  such that for any  $h \leq h_0$ ,  $f_h(\xi,\alpha)$  has a unique branch of solutions in  $U$ ; the branch has the form  $\xi = \xi_h(\alpha)$  and we have for  $0 \leq m \leq r - 1$

$$(4.7) \quad \left| \frac{d^m}{d\alpha^m} (\xi_h(\alpha) - \xi(\alpha)) \right| \leq c \sum_{\ell=0}^m \left| \frac{d^\ell}{d\alpha^\ell} f_h(\xi(\alpha), \alpha) \right|$$

where  $\xi = \xi(a)$  is the (unique) branch of solutions of  $f(\xi, a) = 0$ . Moreover if  $r \geq 2$  and  $(0,0)$  is a nondegenerated turning point for  $f$  then  $f_h$  has a unique nondegenerated turning point  $(\xi_h^0, a_h^0)$  in  $U$  and we have:

$$(4.8) \quad |a_h^0| \leq c \left( |f_h(0,0)| + \left| \frac{\partial f_h}{\partial a}(0,0) \right| \right),$$

$$(4.9) \quad |\xi_h^0| \leq c \left( |f_h(0,0)| + \left| \frac{\partial f_h}{\partial a}(0,0) \right|^2 \right).$$

Remark. In many applications, as we shall see later on,  $|f_h(0,0)|$  is itself of the order of  $\left| \frac{\partial f_h}{\partial a}(0,0) \right|^2$ , which justifies the notations in (4.8), (4.9).

4.2. We shall now assume that  $(0,0)$  is a simple critical point for  $f(\xi, a)$ , that is:  $\frac{\partial f}{\partial \xi}(0,0) = 0$  and the Hessian matrix  $D^2 f^0$  at the origin is nonsingular. It is easy to see that if  $\det(D^2 f^0) > 0$  then the set of solutions of  $f(\xi, a) = 0$  near the origin consists in the isolated point  $(0,0)$ . On the other hand, assume that  $\det(D^2 f^0) < 0$  and consider the auxiliary function

$$(4.10) \quad F(t, \sigma, a) = (t^{-2} f(t\sigma, ta), \sigma^2 + a^2 - 1);$$

It is clear that if we find solutions of  $F = 0$  of the form  $t, \sigma(t), a(t)$  then

$$\begin{cases} \xi(t) = t\sigma(t) \\ \alpha(t) = t\alpha(t) \end{cases}$$

has to be a branch of solutions for  $f(\xi, \alpha) = 0$ . In order to solve  $F = 0$  with the implicit function theorem we look for pairs  $(\sigma_0, \alpha_0)$  such that

$$(4.11) \quad F(0, \sigma_0, \alpha_0) = 0$$

$$(4.12) \quad \mathbb{F}_{(\xi, \alpha)} F(0, \sigma_0, \alpha_0) \text{ is nonsingular.}$$

An easy computation shows that (4.11) represents the intersection of the unit circle with a degenerated hyperbola having vertex in the origin, while (4.12) is satisfied everywhere except on the axes of the same degenerated hyperbola. Hence (4.11)-(4.12) together give four solutions  $(\sigma^1, \alpha^1)$ ,  $(\sigma^2, \alpha^2)$ ,  $(-\sigma^1, -\alpha^1)$ ,  $(-\sigma^2, -\alpha^2)$ . Disregarding the last two for obvious reasons (they will give the same branches with  $-t$  instead of  $t$ ), we are left with two independent solutions. Applying the implicit function theorem we find two branches

$$(4.13) \quad \begin{cases} \xi^i = t\sigma^i(t) & i = 1, 2 \\ \alpha^i = t\alpha^i(t) & i = 1, 2 \end{cases}$$

of solutions of  $f = 0$  crossing transversally at the origin.

Remark. The previous result could have been obtained directly

from Morse lemma (cfr. e.g. [1]). However, an explicit construction of the two branches (4.13) by means of the implicit function theorem will allow the use of Theorem 1 and make the error estimates much easier.

The following lemma will be crucial in the study of the behaviour of the set of solutions of  $f_h(\xi, \alpha) = 0$ .

Lemma 2. Assume that  $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$  at the origin and that  $D^2f$  is nonsingular at the origin. Assume that  $f_h(\xi, \alpha)$  converges uniformly to  $f(\xi, \alpha)$  in  $C^r$ ,  $r \geq 2$ . Then there exist a neighbourhood of the origin,  $N$ , and an  $h_0 > 0$  such that for each  $h < h_0$  there exists a unique point  $(\xi_h^0, \alpha_h^0)$  in  $N$  such that:

$$(4.14) \quad Df_h(\xi_h^0, \alpha_h^0) = 0;$$

moreover we have:

$$(4.15) \quad |\xi_h^0| + |\alpha_h^0| \leq c |Df_h(0,0)|_{\mathbb{R}^2}.$$

The proof is an immediate consequence of Lemma 1.

We consider now the quantity

$$(4.16) \quad K(h) = f_h(\xi_h^0, \alpha_h^0);$$

we remark that  $f_h(\xi, \alpha)$  has a simple critical point (necessarily at  $(\xi_h^0, \alpha_h^0)$ ) iff  $K(h) = 0$ ; in such a case, if

$\det(D_2 f^0) > 0$  the set of zeroes of  $f_h$  consists of the isolated point  $(\xi_h^0, \alpha_h^0)$  and if, on the contrary,  $\det(D_2 f^0) < 0$ , then  $f_h$  has a simple bifurcation point at  $(\xi_h^0, \alpha_h^0)$ .

Roughly speaking, then,  $f_h$  reproduces the behaviour of  $f$  iff  $K(h) = 0$ , which should be regarded as some kind of "miracle"; however, some sufficient conditions to ensure  $K(h) = 0$  in some particular cases (bifurcations from the trivial branch, symmetry breaking bifurcations) can be found in [9], [7].

We set now

$$(4.17) \quad \tilde{f}_h(\xi, \alpha) = f_h(\xi, \alpha) - K(h)$$

and we remark that from (4.15), (4.16) one has

$$(4.18) \quad |K(h)| \leq c(|f_h(0,0)| + |Df_h(0,0)|^2).$$

Introducing the auxiliary function

$$(4.19) \quad F_h(t, \sigma, \alpha) = (t^{-2} \tilde{f}_h(\xi_h^0 + t\sigma, \alpha_h^0 + t\alpha), \sigma^2 + \alpha^2 - 1)$$

comparing with (4.10) and using Theorem 1 we easily get (for  $\det(D^2 f^0) < 0$ ) the following result: the set of zeroes of  $\tilde{f}_h$  is composed of two smooth branches  $(\tilde{\xi}_h^i(t), \tilde{\alpha}_h^i(t))$ ,  $|t| \leq t_0$ , ( $i=1,2$ ) crossing at  $(\xi_h^0, \alpha_h^0)$ ; moreover

$$(4.20) \quad \leq c \left\{ \sum_{\ell=0}^{m+1} \left| \frac{d^\ell}{dt^\ell} f_h(\xi^i(t), \alpha^i(t)) \right| + |f_h(0,0)| + |Df_h(0,0)| \right\}, \quad (0 \leq m \leq r-3, \quad |t| \leq t_0)$$

(see [7] for a detailed proof). Clearly we need, in this case,  $r \geq 3$ ; the interest of (4.20) is mainly in the case  $K(h) = 0$ ; otherwise it will be enough to remark that, from (4.20) (for  $m=0$ ) one has:

$$(4.21) \quad D(S, \tilde{S}_h) \leq c \sup_{|t| \leq t_0} \left( \sum_{i=1}^2 \sum_{\ell=0}^1 \left| \frac{d^\ell}{dt^\ell} f_h(\xi^i(t), \alpha^i(t)) \right| \right)$$

where  $S$  and  $\tilde{S}_h$  are the sets of zeroes of  $f$  and  $\tilde{f}_h$  (respectively) in a fixed ball centered at the origin and the distance  $D(A, B)$  between two closed sets  $A$  and  $B$  is intended as

$$(4.22) \quad D(A, B) = \max(\sup_{x \in A} \inf_{y \in B} \|x-y\|, \sup_{y \in B} \inf_{x \in A} \|x-y\|).$$

We have now to evaluate the distance between  $\tilde{S}_h$  and  $S_h = \text{set of zeroes of } f_h$  in the given ball. For this we remark that since  $\tilde{f}_h$  has a simple critical point we may apply, for any given  $h \leq h_0$ , the Morse lemma to  $\tilde{f}_h$  getting new variables

$$(4.23) \quad (\bar{\xi}, \bar{a}) = R_h(\xi, a)$$

in which

$$(4.24) \quad \tilde{f}_h(\xi, a) = D^2 f_h(\xi_h^0, a_h^0)((\bar{\xi}, \bar{a}), (\bar{\xi}, \bar{a}))$$

(that is,  $\tilde{f}_h$  is a homogeneous polynomial of degree 2 in the variables  $\bar{\xi}$ ,  $\bar{a}$ ). Hence we also have

$$(4.25) \quad f_h(\xi, a) = D^2 f_h(\xi_h^0, a_h^0)((\bar{\xi}, \bar{a}), (\bar{\xi}, \bar{a})) + K(h)$$

and the set  $\bar{S}_h$  of zeroes of  $f_h$  in the  $(\bar{\xi}, \bar{a})$ -plane is as follows:

- a) if  $\det(D^2 f_h(\xi_h^0, a_h^0)) > 0$   $\bar{S}_h$  is an ellipse or the empty set following the sign of  $K(h)$
- b) if  $\det(D^2 f_h(\xi_h^0, a_h^0)) < 0$ .  $\bar{S}_h$  is a nondegenerated hyperbola.

In both cases one can check that

$$(4.26) \quad d(\tilde{S}_h, \bar{S}_h) \leq c |K(h)|^{1/2}$$

provided that  $\bar{S}_h$  is nonempty. Since  $R_h$  in (4.23) is uniformly invertible one has from (4.26)

$$(4.27) \quad d(\tilde{S}_h, S_h) \leq c |K(h)|^{1/2}$$

provided that  $S_h$  is nonempty.

We summarize the previous results in the following theorem.

Theorem 5. Assume that  $f(\xi, a)$  is a  $C^r$  mapping  $\mathbb{R}^2 \rightarrow \mathbb{R}$  with  $r \geq 3$ , and assume that  $f_h(\xi, a)$  converges uniformly to  $f(\xi, a)$  in  $C^r$  in a neighbourhood of the origin. Assume moreover that  $(0, 0)$  is a simple critical point of  $f(\xi, a)$ , that is  $f^0 = f_\xi^0 = f_a^0 = 0$  and  $\det(D^2 f^0) \neq 0$ . Then there exists a neighbourhood of the origin  $N$  and an  $h_0 > 0$  such that for any  $h < h_0$  we have the following results:

i) if  $\det(D^2 f^0) > 0$ ,  $s_h$  is an isola, an isolated point or the empty set following  $K(h) \gtrless 0$ . If  $s_h$  is non-empty:

$$(4.28) \quad \sup_{P \in s_h} |P| \leq c |K(h)|^{1/2};$$

ii) if  $\det(D^2 f^0) < 0$ ,  $s_h$  is diffeomorphic to a hyperboloid (degenerated if  $K(h) = 0$ ) and we have:

$$(4.29) \quad \begin{aligned} & d(s, s_h) \\ & \leq c \left\{ |K(h)|^{1/2} + \sup_{|t| \leq t_0} \sum_{i=1}^2 \sum_{\ell=0}^1 \left| \frac{d^\ell}{dt^\ell} f_h(\xi^i(t), a^i(t)) \right| \right\} \end{aligned}$$

where  $\xi^i(t)$ ,  $a^i(t)$  ( $|t| \leq t_0$ ,  $i=1,2$ ) are the two branches of solutions of  $f_h = 0$  in  $N$  (and hence  $\frac{d^\ell}{dt^\ell} f_h(\xi^i(t), a^i(t)) \neq 0$ ); moreover, if  $K = 0$  the two branches of solutions of  $f_h = 0$  can be parametrized in such a way that (4.29) holds.

We recall that  $s$  and  $s_h$  are the sets of zeroes of  $f$

and  $f_h$  (respectively) in  $N$  and that  $K(h)$  is defined by (4.14)–(4.16) and bounded by

$$(4.30) \quad |K(h)| \leq c(|f_h(0,0)| + |\partial f_h(0,0)|^2).$$

Remark. One can show (cfr [1]) that in the "pure Galerkin case" described in Section 1.2 one has, with the notations of Section 3.2,

$$(4.31) \quad |f_h(0,0)| \leq (\inf_{v \in V_h} \|u_0 - v_h\|_V + \inf_{v \in V_h} \|T^* \phi_0^* - v_h\|_V)^2$$

where  $T^* \in L(V^*, V)$  is the dual operator of  $T$ . Hence under reasonable compatibility assumptions  $f_h(0,0)$  goes to zero twice as fast as any other term in the previous abstract estimates. In other applications to mixed and hybrid elements a relationship as simple as (4.31) does not hold, but still one can prove that  $f_h(0,0)$  goes to zero with a higher order (usually, a double order); see for instance [6], [7], [3].

## C H A P T E R 5

5.1. We assumed, until now, that our problem was governed just by a real parameter  $\lambda \in \mathbb{R}$ . In fact, on one hand, many physical problems are actually governed by more than one parameter; on the other hand, other parameters could be considered, from the theoretical point of view, as "imperfection parameters" in order to see if, in some "expanded space" the numerical discretization reproduces the whole bifurcation diagram. We shall give a simple example in order to make our statements more clear. Assume that

$$(5.1) \quad f(\xi, \alpha) = \xi^2 + \alpha^2$$

and that  $f_h(\xi, \alpha)$  is a  $C^\infty$  function which converges uniformly to  $f(\xi, \alpha)$  with all the derivatives. As we have seen, the set of solutions of  $f_h(\xi, \alpha) = 0$  can be: 1) an isolated point; 2) an isola; 3) the empty set. There is little doubt that, from the qualitative point of view, the way in which the solution set  $S_h$  of  $f_h(\xi, \alpha) = 0$  reproduces the solution set  $S$  of  $f(\xi, \alpha) = 0$  is quite unsatisfactory. Assume now that, instead, we have a two parameter problem

$$(5.2) \quad f(\xi, \alpha, \mu) = 0$$

with  $f(\xi, \alpha, \mu) = \xi^2 + \alpha^2 + \mu$ . For the moment we may assume that  $\mu$  is another parameter "controlled from the exterior" or that  $\mu$  is an "imperfection parameter." Suppose now

that we are given a family  $f_h(\xi, \alpha, \mu)$  of  $C^\infty$ -functions which converges to  $f(\xi, \alpha, \mu)$  uniformly with all the derivatives. Then since  $\frac{\partial f}{\partial \mu}(0, 0, 0) \neq 0$  the implicit function theorem and Theorem 1 give existence and uniqueness of the functions

$$(5.3) \quad \mu = \mu(\xi, \alpha) \quad (\text{actually } \mu = -\alpha^2 - \xi^2)$$

$$(5.4) \quad \alpha_h = \mu_h(\xi, \alpha)$$

such that  $f(\xi, \alpha, \mu(\xi, \alpha))$  and  $f_h(\xi, \alpha, \mu_h(\xi, \alpha))$  vanish identically in a neighbourhood of  $(0, 0)$ . Moreover,

$$(5.5) \quad |D^m(\mu_h(\xi, \alpha) - \mu(\xi, \alpha))| \leq c_m \sum_{\ell=0}^m |D^\ell f_h(\xi, \alpha, \mu(\xi, \alpha))|.$$

Since  $D\mu(0, 0) = (\frac{\partial \mu}{\partial \xi}(0, 0), \frac{\partial \mu}{\partial \alpha}(0, 0)) = 0$  and  $D(D\mu) = D^2\mu$  is nonsingular, Lemma 1 ensures the existence of a unique point  $(\xi_h^0, \alpha_h^0)$  where  $D\mu_h$  vanishes. Setting  $\mu_h^0 = \mu_h(\xi_h^0, \alpha_h^0)$  it is easy to check that  $\mu_h^0$  is an absolute maximum for  $\mu_h$  in a neighbourhood of the origin so that the solution set

$S_h(\bar{\mu})$  of  $f_h(\xi, \alpha, \bar{\mu})$  in the  $(\xi, \alpha)$ -plane is empty for  $\bar{\mu} > \mu_h^0$ , is reduced to the isolated point  $(\xi_h^0, \alpha_h^0)$  for  $\bar{\mu} = \mu_h^0$  and is an isolas for  $\bar{\mu} < \mu_h^0$ . Hence, if we consider  $\mu$  as an "imperfection parameter" we may draw the following pictures, for the continuous and for the discretized problem, in the "space of perturbations"  $\mu \in \mathbb{R}$

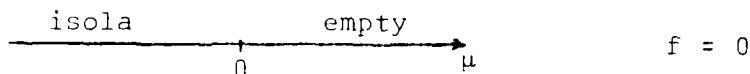


Fig. 5.1

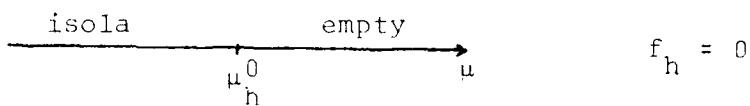


Fig. 5.2

and the qualitative reproduction of the bifurcation diagram is by far more satisfactory. We may also prove, with little effort, that

$$(5.6) \quad |\mu_h^0| \leq c \left( |f_h(0,0,0)| + \left| \frac{\partial f_h}{\partial \alpha}(0,0,0) \right|^2 + \left| \frac{\partial f_h}{\partial \xi}(0,0,0) \right|^2 \right)$$

which, in the applications, may provide, as we have seen, a better order of convergence. Suppose now that we consider  $\mu$  as a parameter controlled by the exterior. An elementary computation shows that in the "space of parameters"  $(\xi, \mu) \in \mathbb{R}^2$  we have for  $f(\xi, \alpha, \mu) = 0$  the following situation

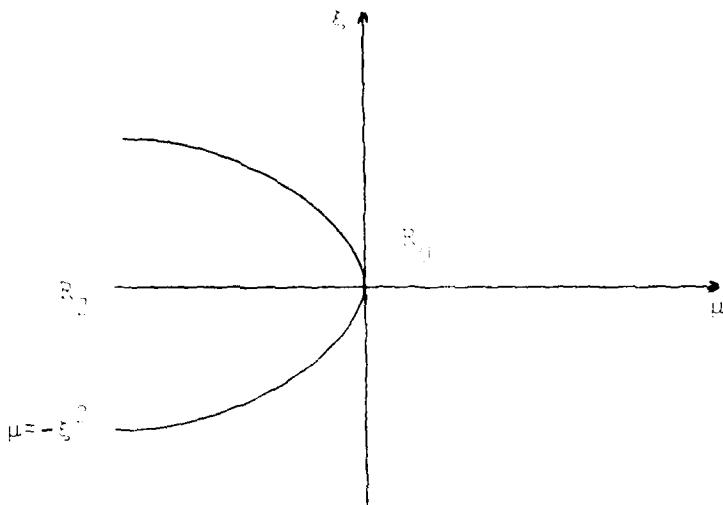


Fig. 5.7

where the parabola  $\mu = -\xi^2$  separates the two regions

$R_0 = \{(\bar{\mu}, \bar{\xi}) \mid f(\bar{\xi}, \alpha, \bar{\mu}) = 0 \text{ has no solutions}\}$  and

$R_2 = \{(\bar{\mu}, \bar{\xi}) \mid f(\bar{\xi}, \alpha, \bar{\mu}) = 0 \text{ has two solutions } \alpha_1 \neq \alpha_2\}$ . Obviously a "double" solution is present on  $\mu = -\xi^2$ . Let us consider now the mapping

$$(5.7) \quad G(\xi, \alpha, \mu) = (f(\xi, \alpha, \mu), \frac{\partial f}{\partial \alpha}(\xi, \alpha, \mu))$$

and its approximation

$$(5.8) \quad G_h(\xi, \alpha, \mu) = (f_h(\xi, \alpha, \mu), \frac{\partial f_h}{\partial \alpha}(\xi, \alpha, \mu))$$

Clearly  $G(0, 0, 0) = 0$  and  $D_{(\alpha, \mu)}G(0, 0, 0) = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$  = non-singular. Hence we may have  $\alpha = \alpha(\xi)$ ,  $\mu = \tilde{\mu}(\xi)$  as implicit

functions satisfying

$$(5.9) \quad G(\xi, \alpha(\xi), \tilde{\mu}(\xi)) = 0.$$

Theorem 1 yields now  $\alpha_h(\xi)$ ,  $\tilde{\mu}_h(\xi)$  such that

$$(5.10) \quad G_h(\xi, \alpha_h(\xi), \tilde{\mu}_h(\xi)) = 0$$

with an estimate on  $\alpha_h(\xi) - \alpha(\xi)$  and  $\tilde{\mu}_h(\xi) - \tilde{\mu}(\xi)$ . Clearly

$$(5.11) \quad \tilde{\mu}_h(\xi) = \mu_h(\xi, \alpha_h(\xi))$$

from the uniqueness of the implicit function; it is also obvious that  $\tilde{\mu}(\xi) = -\xi^2$ . Therefore we have, in the space of parameters  $(\xi, \mu)$  the following picture for the approximate problem

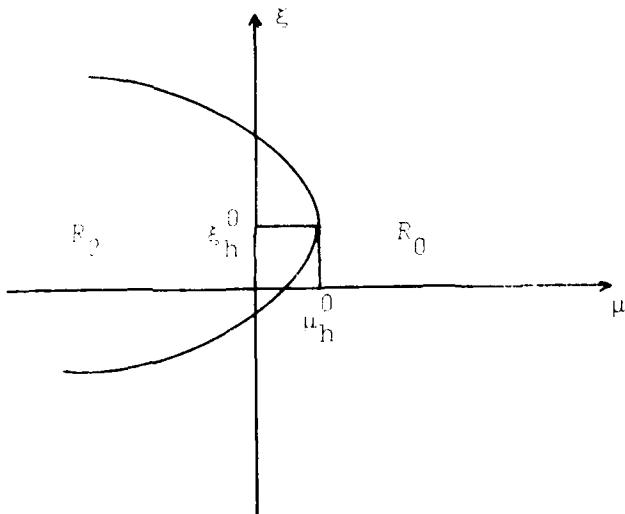


Fig. 5.4

It is also easy to show that  $\frac{\partial^2 \tilde{\mu}_h}{\partial \xi^2} < 0$  near zero (it converges uniformly to  $-2 \equiv \frac{\partial^2 \tilde{\mu}}{\partial \xi^2}$ ) and that there are no regions, in some fixed neighbourhood of the origin, where  $f_h(\xi, \alpha, \tilde{\mu}) = 0$  has more than two solutions in  $\alpha$ . Again a comparison between figures 5.3 and 5.4 shows now a qualitative agreement which is much more satisfactory. We can summarize the results obtained on that simple example in the following theorem.

Theorem 6. Assume that

$$(5.12) \quad f(\xi, \alpha, \mu) = \xi^2 + \alpha^2 + \mu$$

and assume that we are given a family  $f_h(\xi, \alpha, \mu)$  of  $C^\infty$  functions which converges to  $f(\xi, \alpha, \mu)$  uniformly with all the derivatives in a neighbourhood of the origin. Then we have the following results:

a) In the whole space  $(\xi, \alpha, \mu)$ : there exists, for  $h < h_0$ , a unique  $C^\infty$  mapping

$$(5.13) \quad \mu = \mu_h(\xi, \alpha)$$

in a neighbourhood of  $(0, 0)$  such that

$$(5.14) \quad f_h(\xi, \alpha, \mu_h(\xi, \alpha)) = 0$$

and we have:

$$(5.15) \quad |D^m(\mu_h(\xi, \alpha) - \mu(\xi, \alpha))| \leq c_m \sum_{\ell=0}^m |D^\ell f_h(\xi, \alpha, \mu(\xi, \alpha))|$$

where  $\mu(\xi, \alpha) = -\xi^2 - \alpha^2$ .

b) In the "space of perturbations"  $\mu \in \mathbb{R}$ : there exists a unique  $\mu_h^0$  such that the set of solution of  $f_h(\xi, \alpha, \bar{\mu}) = 0$  in the  $(\xi, \alpha)$  plane is: i) an isolator for  $\bar{\mu} < \mu_h^0$ ; ii) an isolated point for  $\bar{\mu} = \mu_h^0$ ; iii) empty for  $\bar{\mu} > \mu_h^0$ . Moreover

$$(5.16) \quad |\mu_h^0| \leq c(|f_h(0, 0, 0)| + |D_{(\xi, \alpha)} f_h(0, 0, 0)|^2).$$

c) In the "space of parameters"  $(\xi, \alpha) \in \mathbb{R}^2$ : there exists a unique mapping

$$(5.17) \quad \mu = \tilde{\mu}_h(\xi)$$

which divides the  $(\xi, \mu)$  plane in two regions  $R_0^h$  and  $R_2^h$ , such that for  $(\xi, \bar{\mu}) \in R_0^h$  the equation (in  $\alpha$ )  $f_h(\xi, \alpha, \bar{\mu}) = 0$  has no solutions and for  $(\xi, \bar{\mu}) \in R_2^h$  the equation (in  $\alpha$ )  $f_h(\xi, \alpha, \bar{\mu}) = 0$  has two distinct solutions. Moreover we have:

$$(5.18) \quad \left| \frac{d}{d\xi} (\tilde{\mu}_h(\xi) - \bar{\mu}(\xi)) \right| \leq c_m \left\{ \sum_{\ell=1}^m \left| \frac{d^\ell}{d\xi^\ell} \tilde{\mu}_h(\xi, 0, \tilde{\mu}(\xi)) \right| + \left| \frac{d}{d\xi} \frac{\partial f_h}{\partial \alpha} (\xi, 0, \tilde{\mu}(\xi)) \right| \right\}.$$

where  $\tilde{\mu}(\xi) = -\xi^2$  realizes the analogous partition of  $(\xi, \mu)$  into  $R_0$  and  $R_2$  for the continuous problem  $f(\xi, \alpha, \mu) = 0$ .

Remark. Our assumption that  $f(\xi, \alpha, \mu)$  has exactly the form (5.1) is obviously unnecessary; as a matter of fact, in the proof of Theorem 6 we only need suitable nondegeneracy conditions on the partial derivatives of  $f$  at the origin.

Remark. An exchange in roles of  $\mu$  and  $\xi$ , by considering  $\mu$  as a parameter and  $\xi$  as a perturbation will not give interesting results. In fact, the equation  $\alpha^2 + \mu = 0$  has a nondegenerated turning point with respect to the parameter  $\mu$  and such a diagram, as we have seen, is already stable under small perturbations (see Sect. 4.1).

5.2. Let us consider now a different example; assume that  $f(\xi, \alpha, \mu)$  has the form

$$(5.19) \quad f(\xi, \alpha, \mu) = \alpha^3 - \mu\alpha + \xi$$

(this form is typical, for instance, in the von Kármán nonlinear plate bending equations). The form (5.19) can be considered, on one hand, as a perturbation of the "pitchfork" form

$$(5.20) \quad f(\xi, \alpha) = \alpha^3 - \mu\alpha = 0$$

or, on the other hand, as a perturbation of the "nondegenerated"

hysteresis" form

$$(5.21) \quad f(\xi, \alpha) = \alpha^3 - \xi = 0.$$

Let now assume that we are given a family  $f_h(\xi, \alpha, \mu)$  of  $C^\infty$  functions which converges to  $f(\xi, \alpha, \mu)$  uniformly with all the derivatives in a neighbourhood of the origin, when  $h$  tends to zero. It is easy to check that for  $h \leq h_0$  there exists  $\xi = \xi_h(\alpha, \mu)$  such that

$$(5.22) \quad f_h(\xi_h(\alpha, \mu), \alpha, \mu) \equiv 0$$

and  $\xi_h(\alpha, \mu)$  converges to  $\xi(\alpha, \mu) = \mu\alpha - \alpha^3$  with all the derivatives. This solves somehow the problem "in the whole space." Let us now have a look to the "space of parameters"  $(\xi, \mu)$ . It is easy to see that, for the continuous problem  $f(\xi, \alpha, \mu) = 0$  the curve

$$(5.23) \quad \begin{cases} \mu = 3t^2 \\ \xi = 2t^3 \end{cases}$$

separates the two regions  $R_1$  and  $R_3$  of "one solution" and "three solutions" (respectively) as in Fig 5.5.

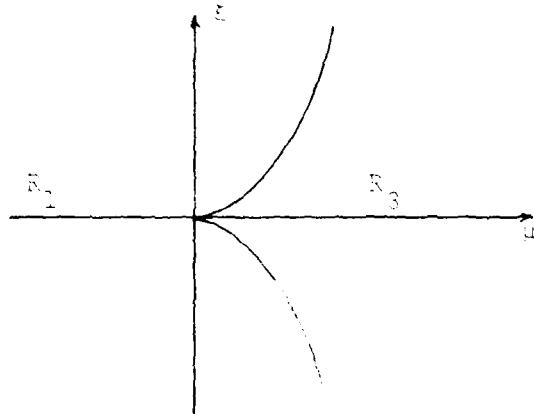


Fig. 5.3

Consider now the mapping:

$$(5.24) \quad G(\xi, \alpha, \mu) = \left( f(\xi, \alpha, \mu), \frac{\partial f}{\partial \xi}(\xi, \alpha, \mu) \right)$$

$$(5.25) \quad G(\xi, \alpha, \mu) = \left( f_h(\xi, \alpha, \mu), \frac{\partial f_h}{\partial \xi}(\xi, \alpha, \mu) \right).$$

Clearly  $G(0, 0, 0) = (0, 0)$  and  $D_{(\xi, \mu)}G(0, 0, 0)$  is nonsingular; hence, from Theorem 1, there exist, together with the implicit functions  $\xi = 2\alpha^3$ ,  $\mu = 3\alpha^2$ , two "discrete implicit functions"

$$(5.26) \quad \tilde{\xi}_h(\alpha), \quad \tilde{\mu}_h(\alpha)$$

such that

$$(5.27) \quad f_h(\tilde{\xi}_h(\alpha), \alpha, \tilde{\mu}_h(\alpha)) = 0;$$

$$(5.28) \quad \frac{\partial f_h}{\partial \alpha} (\tilde{\xi}_h(\alpha), \alpha, \tilde{\mu}_h(\alpha)) \equiv 0.$$

The two functions (5.26) define parametrically a curve which converge to  $\mu = 3 \left( \sqrt[3]{\frac{\xi}{2}} \right)^2$ . We may remark that (5.27) implies

$$(5.29) \quad \frac{\partial f_h}{\partial \alpha} + \frac{\partial f_h}{\partial \xi} \frac{\partial \tilde{\xi}_h}{\partial \alpha} + \frac{\partial f_h}{\partial \mu} \frac{\partial \tilde{\mu}_h}{\partial \alpha} \equiv 0$$

and since  $\frac{\partial f_h}{\partial \xi} \neq 0$  one gets

$$(5.30) \quad \frac{\partial \tilde{\mu}_h}{\partial \alpha} = 0 \implies \frac{\partial \tilde{\xi}_h}{\partial \alpha} = 0$$

so that the curve  $\mu = \tilde{\mu}_h(\alpha)$ ,  $\xi = \tilde{\xi}_h(\alpha)$  does actually have a cusp, because  $\tilde{\mu}_h(\alpha)$  converges uniformly to  $3\alpha^2$  with all the derivatives and hence  $\frac{\partial \tilde{\mu}_h}{\partial \alpha}$  has one and only one zero (cfr. Lemma 1). Therefore, in the "space of parameters" the behaviour of the discrete problem is of the type of figure 5.6.

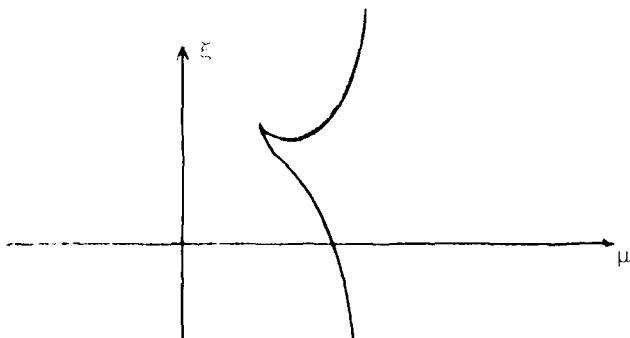


Fig. 5.6

We may now pass to the "space of perturbations" in two different ways: by considering  $\mu$  as a perturbation or by considering  $\xi$  as a perturbation; the situation is different in the two cases: if  $\mu$  is a perturbation we have diagrams of the following type

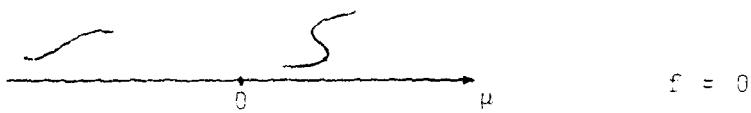


Fig. 5.7

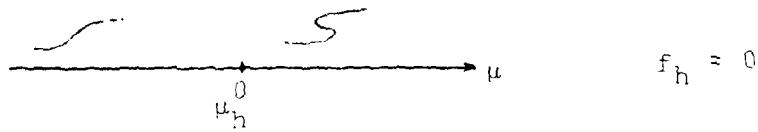


Fig. 5.8

If  $\mu_h^0$  being the abscissa of the cusp in Fig. 5.6,  $\xi$  is considered as a perturbation parameter we may have



Fig. 5.9

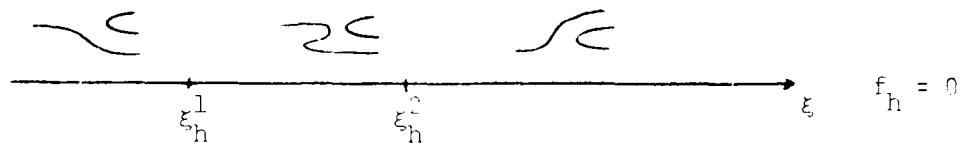


Fig. 5.10

where  $\xi_h^2$  will correspond to the presence of a hysteresis point and  $\xi_h^1$  to the presence of a bifurcation point. We can say, therefore, that the introduction of the perturbation parameter  $\mu$  has clarified the nondegenerated hysteresis case (5.11) while the introduction of the perturbation parameter  $\xi$  has not resolved in a satisfactory way the pitchfork case (5.20).

Remark. The error estimates, for this case, similar to the ones of Theorem 5 are not difficult to work out; we leave it as an exercise.

5.3. It seems now natural to ask the following question: what is the minimum number of "perturbation parameters," for a given problem, that will ensure, for the discrete problem, a correct approximation of the whole bifurcation diagram in the "space of perturbations"? Unfortunately we are not able to answer such a question. However, our guess is that the required number should be the number of parameters of the minimal universal unfolding of the original problem, in the

sense of [10]. For instance, such number is 1 for problem (5.1), 1 for the problem (5.21) and 2 for the pitchfork case (5.20), the additional perturbation being of type

$$(5.32) \quad f(\xi_1, \xi_2, x, u) = u^3 + \alpha u + \xi_1 + x^2 \xi_2.$$

In such case it is possible to show that in the  $(\xi_1, \xi_2)$  plane the two curves

$$(5.33) \quad \xi_1 = 0, \quad \xi_2 = \frac{x^2}{\alpha}$$

separate the regions  $1 \rightarrow 3$  from the regions  $1 \rightarrow 3 \rightarrow 1 \rightarrow 1$  (see Fig. 5.11).

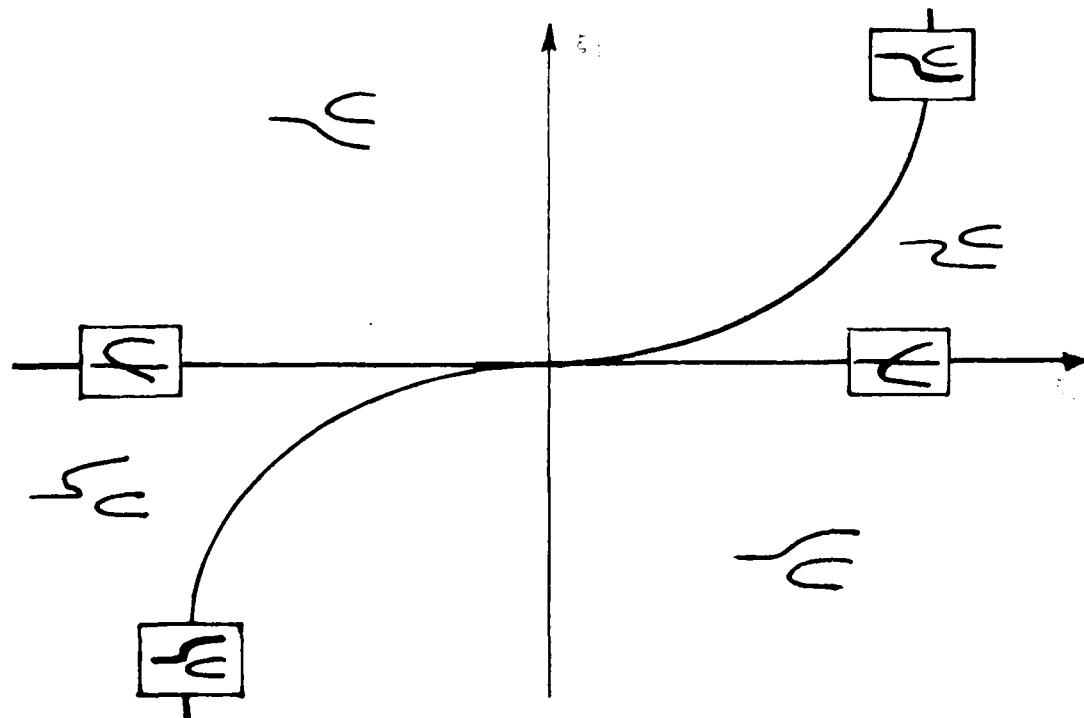


Figure 5.11

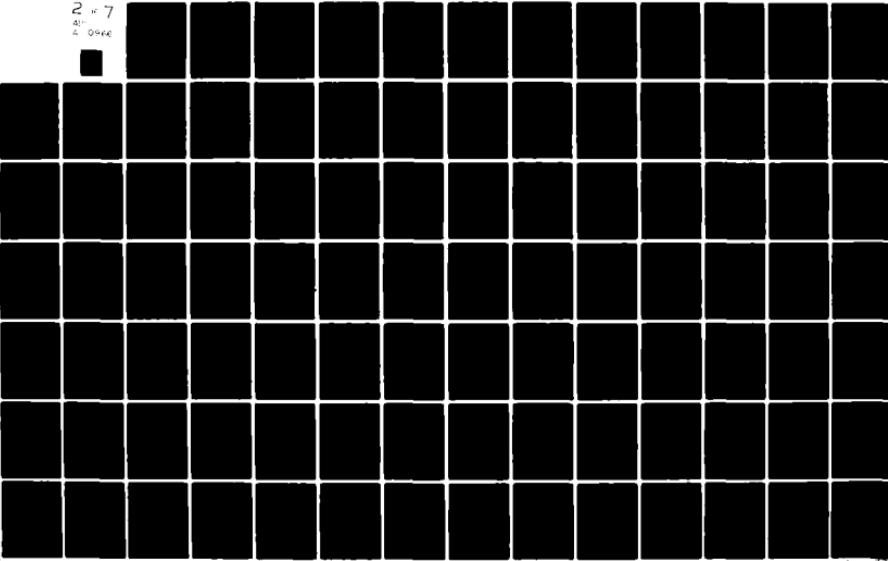
AD-A110 966 MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS F/G 12/1  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

UNCLASSIFIED

AFOSR-TR-82-0047

NL

2 x 7  
20  
2 0460



It is possible to show (cfr. [4]) that two similar curves exist for the discrete problem, that they cross at just one point (pitchfork for the discrete problem) and to estimate their speed of convergence. For other examples and different applications, see [4].

## REFERENCES

- [1] Berger, M. S., Nonlinearity and functional analysis, Acad. Press. New York, 1971.
- [2] Ciarlet, P. G., The finite element method for elliptic problems, North-Holland, Amsterdam, 1978.
- [3] Brezzi, F., "Hibrid approximations of nonlinear plate bending problems," Proc. of "International Symposium on Hybrid and Mixed F.E.M." Atlanta, 1981. To appear.
- [4] Brezzi, F. and Fujii, H., Approximation of nonlinear problems with more then one parameters." To appear.
- [5] Brezzi, F., Rappaz, J. and Raviart, P.-A., "Finite dimensional approximation of nonlinear problems. Part I, branches of nonsingular solutions," Num Math. 36 (1980), 1-25.
- [6] \_\_\_\_\_, "Finite dimensional approximation of nonlinear problems. Part II, limit points." To appear on Num. Math.
- [7] \_\_\_\_\_, "Finite dimensional approximation of nonlinear problems. Part III, simple bifurcation points." Submitted to Num. Math.
- [8] Fujii, H, Mimura, M. and Nishiura, Y., "A picture of Global Bifurcation Diagram in Ecological Interacting and Diffusing Systems," Res. Rep. KSU-ICS 70-11 (1979). Kyoto Sangyo University.
- [9] Fujii, H. and Yamaguti, M., "Structure of singularities and its numerical realization in nonlinear elasticity," J. Math. Kyoto University 20 (1980), 489-590.
- [10] Golubitsky, M. and Schaeffer, D., "A theory for Imperfect Bifurcation via Singularity Theory," Comm. Pure Appl. Math. 32 (1979), 21-98.
- [11] Keller, H. B., "Numerical solution of bifurcation and nonlinear eigenvalue problems," in "Applications of bifurcation Theory" (P. H. Rabinowitz, ed.), Acad. Press, New York, 1978.
- [12] \_\_\_\_\_, "Two new bifurcation fenomena," Rapp. 369 (Nov. 1979) I.N.R.I.A. (Laboria) Le Chesnay (France).

- [13] Kikuchi, F., "An iterative finite element scheme for bifurcation analysis of semi-linear elliptic equations," Report Inst. Space Aero. Sci., 542 (1976), Univ. of Tokyo.
- [14] \_\_\_\_\_, "Finite element approximations to bifurcation problems of turning point type," Theoretical and Applied Mechanics 27 (1979), 99-114.
- [15] Rappaz, J. and Raugel, G., "Finite dimensional approximation of bifurcation problems at a multiple eigenvalue." To appear.

# TWO-DIMENSIONAL APPROXIMATIONS OF THREE-DIMENSIONAL MODELS IN NONLINEAR PLATE THEORY<sup>(\*)</sup>

Philippe G. CIARLET\*

## *Abstract*

The asymptotic expansion method, with the thickness as the parameter, is applied to the nonlinear, three-dimensional, equations for the equilibrium of elastic plates under suitable loads and appropriate boundary conditions. It is shown that the leading term of the expansion is solution of a system of equations equivalent to a well-known two-dimensional nonlinear plate model, namely the von Kármán equations.

The existence of solutions of the two-dimensional problem is established in all cases (by contrast with the three-dimensional model, where no satisfactory existence theory is as yet available). It is also shown that the displacement and the stress corresponding to the leading term of the expansion have the specific form generally assumed *a priori* in the usual derivations of two-dimensional plate models. In particular, the displacement field is of Kirchhoff-Love type.

This approach clarifies in particular the nature of the admissible three-dimensional boundary conditions for a given two-dimensional plate model. A discussion is also given regarding the class of admissible three-dimensional models.

---

(\*) To appear in the Journal of Elasticity.

\*Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris.

1. INTRODUCTION. This paper gives a brief description of a method for deriving known nonlinear two-dimensional plate models from general nonlinear three-dimensional elasticity models. It is based on, and extends as regards the consideration of more general constitutive equations, Ciarlet [1980], where complete proofs can be found.

Our approach is based on the *asymptotic expansion method*, applied to (nonlinear in the present case) problems posed in variational form. Without any *a priori assumption*, either geometrical or mechanical in nature, it is shown that the first term in the expansion is solution of a *two-dimensional plate model*, equivalent to the von Kármán equations.

A feature of the method is to clearly delineate the type of *boundary conditions* for the three-dimensional model which lead to a specific two-dimensional plate model.

Another aspect of the method is that the displacement and stress components corresponding to the first term in the asymptotic expansion are of the specific forms generally assumed in the literature as a result of appropriate *a priori assumptions*. For instance we shall find that the displacement field is necessarily of Kirchhoff-Love type, while this is generally an *a priori assumption* of a geometrical nature.

In other words, the *asymptotic expansion method* has been shown to apply equally well to :

- (i) *linear plate models* [Ciarlet and Destuynder, 1979a], for which it provides in addition a satisfactory *error analysis* [Destuynder, 1979] between the three-dimensional and two-dimensional solutions (the error analysis rests upon methods developed in Lions [1973]) ;
- (ii) *eigenvalue problems for plates* [Ciarlet and Kesavan, 1980] ;
- (iii) *linear shell models* [Destuynder, 1979].

It is also worth mentioning that whereas the asymptotic expansion method is commonly used for linear problems, it is seldom applied to nonlinear problems ; in this direction, see however Lions [1973], Rigolot [1977].

Let us review some of the notation used in this paper. The usual partial derivatives will be written  $\partial_i v = \partial v / \partial x_i$ ,  $\partial_{ij} v = \partial^2 v / \partial x_i \partial x_j$ , etc ... . If  $\Omega$  is an open subset of  $\mathbb{R}^n$ , we denote by  $W^{m,p}(\Omega)$ ,  $m \in \mathbb{N}$ ,  $1 \leq p$ , or  $H^m(\Omega)$  if  $p = 2$ , the standard Sobolev spaces.

We shall omit the symbol  $dx$  in an integral of the form  $\int_X f(x) dx$ , except in those integrals where the variable of integration is  $x_3 \in [-1,1]$ , in which case the specific symbol  $dt$  will be used.

As a rule, Greek indices ;  $\alpha, \beta, \mu, \dots$ , take their values in the set  $\{1, 2\}$ , while Latin indices :  $i, j, p, \dots$ , take their values in the set  $\{1, 2, 3\}$ . The repeated index convention for summation is also systematically used, in conjunction with the above rule.

With each vector-valued function  $v = (v_i) : \Omega \subset \mathbb{R}^3 \rightarrow \mathbb{R}^3$ , thought of as being a displacement field in  $\mathbb{R}^3$ , we associate the symmetric tensors  $\gamma(v) = (\gamma_{ij}(v))$  and  $\bar{\gamma}(v) = (\bar{\gamma}_{ij}(v)) : \Omega \subset \mathbb{R}^3 \rightarrow \mathbb{R}^9$  respectively defined by

$$\gamma_{ij}(v) = \frac{1}{2}(\partial_i v_j + \partial_j v_i),$$

$$\bar{\gamma}_{ij}(v) = \gamma_{ij}(v) + \frac{1}{2}\partial_i v_k \partial_j v_k,$$

which are the *linearized strain tensor*, and the *strain tensor*, respectively.

Finally, if  $C$  is a square matrix, we denote by  $\text{tr}(C)$  and  $\det(C)$  its trace and its determinant, respectively.

2. THE THREE-DIMENSIONAL MODEL. Let  $(e_i)$  be an orthonormal basis in  $\mathbb{R}^3$ , and let  $w$  be a bounded open subset of the plane spanned by  $(e_\alpha)$ , with a sufficiently smooth boundary  $\gamma$ . Given a constant  $\epsilon > 0$ , we let

$$\Omega^\epsilon = w \times [-\epsilon, \epsilon], \quad \Gamma_0^\epsilon = \gamma \times [-\epsilon, \epsilon],$$

$$\Gamma_+^\epsilon = w \times \{\epsilon\}, \quad \Gamma_-^\epsilon = w \times \{-\epsilon\},$$

so that the boundary of the open subset  $\Omega^\epsilon$  of  $\mathbb{R}^3$  is partitioned into the lateral surface  $\Gamma_0^\epsilon$  and the upper and lower faces  $\Gamma_+^\epsilon$  and  $\Gamma_-^\epsilon$ .

The problem consists in finding the displacement vector field

$u = (u_i) : \bar{\Omega} \rightarrow \mathbb{R}^3$  and the second Piola-Kirchhoff tensor field  $\sigma = (\sigma_{ij}) : \bar{\Omega} \rightarrow \mathbb{R}^{9 \times 9}$  of a three-dimensional body which occupies the set  $\bar{\Omega}$  in the absence of applied forces. Because the thickness  $2\epsilon$  of the body is considered to be "small" compared to the dimensions of the set  $w$ , the body is called a plate, with middle surface  $w$ .

The plate is subjected to three kinds of given forces :

(i) Body forces throughout  $\Omega^\epsilon$ , of density

$$(f_i^\epsilon) = (0, 0, f_3^\epsilon);$$

(ii) Superficial forces on the upper and lower faces  $\Gamma_+^\epsilon$  and  $\Gamma_-^\epsilon$ , of density

$$(g_i^\epsilon) = (0, 0, g_3^\epsilon);$$

(iii) Superficial forces along the lateral surface  $\Gamma_0^\epsilon$ , of which only the resulting density

$$(h_i^\epsilon) = (h_1^\epsilon, h_2^\epsilon, 0),$$

i.e., after integration across the thickness of the plate (cf. (2.4) below), is known along the boundary  $\gamma$  of the middle surface  $w$  (as a consequence, the functions  $h_\alpha^\epsilon$  are given only on  $\gamma$ ).

As regards the boundary conditions involving the displacement field  $(u_i)$ , we assume that :

$$\left. \begin{array}{l} u_1 \text{ and } u_2 \text{ are independent of } x_3, \\ u_3 = 0, \end{array} \right\} \text{on } \Gamma_0^\epsilon.$$

These conditions are readily verified to be complementary to those involving the functions  $h_\alpha^\epsilon$  in the variational formulation of the problem (cf. (2.20) below).

Following Truesdell and Noll [1965], or Wang and Truesdell [1973], the associated equations of finite elastostatics, which express the elastic equilibrium of the plate, take the following form :

$$(2.1) \quad - \partial_j (\sigma_{ij} + \sigma_{kj} \partial_k u_i) = f_i^\epsilon \text{ in } \Omega^\epsilon,$$

$$(2.2) \quad \sigma_{ij} = \sigma_{ji} \text{ in } \Omega^\epsilon,$$

$$(2.3) \quad \sigma_{i3} + \sigma_{k3} \partial_k u_i = g_i^\epsilon \text{ on } \Gamma_t^\epsilon,$$

$$(2.4) \quad \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} (\sigma_{\alpha\beta} + \sigma_{k\beta} \partial_k u_\alpha) v_\beta = h_\alpha^\epsilon \text{ on } \gamma,$$

$$(2.5) \quad u_1, u_2 \text{ are independent of } x_3 \text{ on } \Gamma_0^\epsilon,$$

$$(2.6) \quad u_3 = 0 \text{ on } \Gamma_0^\epsilon,$$

where  $v = (v_\alpha)$  denotes the unit outer normal vector along  $\gamma$  (and consequently, also along the lateral surface  $\Gamma_0^\epsilon$ ).

*Remark 2.1.* The reason why we set  $f_1^\epsilon = f_2^\epsilon = 0$ ,  $g_1^\epsilon = g_2^\epsilon = 0$ , and  $h_3^\epsilon = 0$ , is simply that the consideration of such more general applied forces leads to plate models different from (and more complicated than) the von Kármán equations. ■

*Remark 2.2.* If instead of the boundary conditions (2.4)-(2.6), we had chosen the (perhaps more familiar) boundary conditions :

$$(\sigma_{\alpha\beta} + \sigma_{k\beta} \partial_k u_\alpha) v_\beta = h_\alpha^\epsilon \text{ on } \Gamma_0^\epsilon,$$

$$u_3 = 0 \text{ on } \Gamma_0^\epsilon,$$

serious difficulties would arise in the subsequent analysis. In particular, it seems that this type of boundary conditions along the lateral surface does not naturally give rise to a two-dimensional plate model.

According to the Rivlin-Ericksen theorem (cf. Wang & Truesdell [1973]), the most general constitutive equation for an *elastic, isotropic*, material which satisfies the principle of *frame indifference* is of the form :

$$(2.7) \quad \sigma = \sqrt{III_c} \{ \varphi_0 (I_c, II_c, III_c) C^{-1} + \varphi_1 (I_c, II_c, III_c) I + \varphi_2 (I_c, II_c, III_c) C \},$$

where  $I$  denotes the unit matrix,

$$C = I + 2\bar{\gamma}, \text{ with } \bar{\gamma} = \bar{\gamma}(u),$$

denotes the (right) Cauchy-Green tensor,  $I_c$ ,  $II_c$ ,  $III_c$  denote the three principal invariants of the tensor  $C$  (whose eigenvalues are denoted  $\lambda_1, \lambda_2, \lambda_3$ ) :

$$I_c = \lambda_1 + \lambda_2 + \lambda_3 = C_{ii} = \text{tr}(C),$$

$$II_c = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1 = \frac{1}{2}\{(\text{tr}C)^2 - \text{tr}(C^2)\},$$

$$III_c = \lambda_1\lambda_2\lambda_3 = \det(C) = \frac{1}{6}\{(\text{tr}C)^3 - 3\text{tr}C \text{tr}(C^2) + 2\text{tr}(C^3)\},$$

and finally,  $\varphi_0$ ,  $\varphi_1$  and  $\varphi_2$  are arbitrary functions.

Assuming the functions  $\varphi_0$ ,  $\varphi_1$ ,  $\varphi_2$  to be smooth enough, one can write a Taylor expansion of (2.7) around a natural state ( $\sigma = 0$  for  $C = I$ ) in terms of the strain tensor  $\bar{\gamma}$ . Thus for instance, if we limit ourselves to second order terms, we find a constitutive equation of the form

$$(2.8) \quad \sigma = \lambda(\text{tr}\bar{\gamma})I + 2\mu\bar{\gamma} + a\bar{\gamma}^2 + b(\text{tr}\bar{\gamma})\bar{\gamma} + c(\text{tr}\bar{\gamma})^2I + d(\text{tr}\bar{\gamma}^2)I + \dots,$$

where  $\lambda$ ,  $\mu$ ,  $a$ ,  $b$ ,  $c$ ,  $d$  are constants. The two constants  $\lambda$ ,  $\mu$  are the Lamé coefficients of elasticity ; they satisfy the inequalities (cf. Wang & Truesdell [1973])

$$(2.9) \quad \lambda > 0, \mu > 0.$$

The same type of constitutive equation can also be drawn from the assumption that the material is *hyperelastic*, i.e., that there exists a strain energy function

$$(2.10) \quad W(F) = W(\sigma_1, \sigma_2, \sigma_3),$$

where

$$F = (F_{ij}) = (\partial_j u_i)$$

denotes the deformation gradient matrix, and

$$\sigma_1 = \text{tr} \bar{\gamma} = \bar{\gamma}_{ii},$$

$$\sigma_2 = \text{tr}(\bar{\gamma}^2) = \bar{\gamma}_{ij} \bar{\gamma}_{ji},$$

$$\sigma_3 = \text{tr}(\bar{\gamma}^3) = \bar{\gamma}_{ij} \bar{\gamma}_{jk} \bar{\gamma}_{ki},$$

in such a way that the first Piola-Kirchhoff stress tensor

$$(2.11) \quad t_{ij} \stackrel{\text{def}}{=} \sigma_{ij} + \sigma_{kj} \partial_k u_i$$

satisfies

$$(2.12) \quad t_{ij} = \frac{\partial W}{\partial F_{ij}}.$$

Then if we express that the energy

$$(2.13) \quad J(v) = \int_{\Omega^\epsilon} W(F) - \left( \int_{\Omega^\epsilon} f_3^\epsilon v_3 + \int_{\Gamma_+^\epsilon \cup \Gamma_-^\epsilon} g_3^\epsilon v_3 + \int_{\gamma} \left\{ \int_{-\epsilon}^{\epsilon} v_\alpha dx_3 \right\} h_\alpha^\epsilon \right)$$

is stationary (i.e., its derivative vanishes) when the functions  $v$  span a space of smooth enough functions which satisfy the boundary conditions (2.5)-(2.6), we are naturally led to a constitutive equation. To be more specific, assume that we can expand the strain energy function (2.10) in terms of powers of  $\sigma_1, \sigma_2, \sigma_3$ . Then if we limit ourselves to the quadratic and cubic terms in this expansion, i.e., if we write

$$(2.14) \quad W(\sigma_1, \sigma_2, \sigma_3) = \frac{\lambda}{2} \sigma_1^2 + \mu \sigma_2 + \frac{\Lambda}{3} \sigma_1^3 + B \sigma_1 \sigma_2 + C \sigma_3 + \dots,$$

we find that (cf. John [1971], Novozhilov [1953])

$$(2.15) \sigma = \lambda \sigma_1 I + 2\mu \bar{\gamma} + (C+4\mu) \bar{\gamma}^2 + (2B+2\lambda) \sigma_1 \bar{\gamma} + (A - \frac{\lambda}{2}) \sigma_1^2 I + (B-\mu) \sigma_2 I + \dots .$$

In other words, we find a constitutive equation of the same form as in (2.8), but with only 5 arbitrary constants (instead of 6 in (2.8)), because of the assumption of hyperelasticity.

*Remark 2.3.* When the higher order terms (represented by three dots) are omitted in (2.18), the resulting constitutive equation is sometimes known as *Murnaghan's law*, after Murnaghan [1937], although it seems to have been first considered by Voigt [1893-1894]. For the actual computations of the third order terms in (2.8), see Novozhilov [1953]. ■

We shall henceforth assume that the constitutive equation is a polynomial in terms of the components of the strain tensor  $\bar{\gamma}$ , i.e., we assume that the expansion (2.8) is finite ; hence we do not have to examine questions of convergence in otherwise infinite expansions.

We also make the following assumption, which is crucial for our subsequent purposes, and which shall be commented upon later on (in Section 5) : *The Lamé coefficients appearing in (2.8) are of the form*

$$(2.16) \quad \lambda^\epsilon = \epsilon^{-3} \lambda^1, \quad \mu^\epsilon = \epsilon^{-3} \mu^1,$$

where  $\lambda^1$  and  $\mu^1$  are constants independants of  $\epsilon$ , while the other constants which appear in the constitutive equations (2.8) are independent of  $\epsilon$ .

With each tensor  $X = (X_{ij})$ , we associate the tensor  $Y = (Y_{ij}) = \Lambda X$  defined by ( $\delta_{ij}$  is the Kronecker symbol)

$$Y_{ij} = (\Lambda X)_{ij} = \left(\frac{1+\nu}{E}\right) X_{ij} - \frac{\nu}{E} X_{pp} \delta_{ij},$$

where the constants  $E$  and  $\nu$  are related to the constants  $\lambda^1$ ,  $\mu^1$  appearing in (2.16) by the relations

$$\lambda^1 = \frac{Ev}{(1+v)(1-2v)}, \quad \mu^1 = \frac{E}{2(1+v)}.$$

Since

$$(A^{-1}Y)_{ij} = \lambda^1 Y_{pp} \delta_{ij} + 2\mu^1 Y_{ij},$$

we can also write the constitutive equation (2.8) as

$$(2.17) \quad \epsilon^3 (\Lambda \sigma)_{ij} = \bar{\gamma}_{ij}(u) + \epsilon^3 \sum_{2 \leq q \leq Q} a_{ijk_1 k_2 \dots k_{2q-1} k_{2q}} \bar{\gamma}_{k_1 k_2}(u) \dots \bar{\gamma}_{k_{2q-1} k_{2q}}(u),$$

for appropriate constants  $a_{ijk_1 k_2 \dots k_{2q-1} k_{2q}}$ .

The three-dimensional problem is now completely defined, by the data of the equations of elastic equilibrium (2.1)-(2.6) and of the constitutive equation (2.17).

As regards existence theory for such nonlinear elasticity models, one can extend the analysis given in Ciarlet & Destuynder [1979b] (which relied essentially on the implicit function theorem,  $L^p$ -regularity results for linear elliptic systems, and the fact that the Sobolev spaces  $W^{1,p}(\Omega)$ ,  $\Omega \subset \mathbb{R}^3$ , are Banach algebras for  $p > 3$ ), and show in this fashion that for small enough applied forces  $(f_i) \in (L^p(\Omega))^3$ , the pure Dirichlet problem :

$$-\partial_j (\sigma_{ij} + \sigma_{kj} \partial_k u_i) = f_i \text{ in } \Omega \subset \mathbb{R}^3,$$

$$(\Lambda \sigma)_{ij} = \bar{\gamma}_{ij}(u) + \sum_{2 \leq q \leq Q} a_{ijk_1 k_2 \dots k_{2q-1} k_{2q}} \bar{\gamma}_{k_1 k_2}(u) \dots \bar{\gamma}_{k_{2q-1} k_{2q}}(u),$$

$$u = 0 \text{ on the boundary of } \Omega$$

(assuming the boundary of  $\Omega$  is smooth enough), has a solution in the space  $(W_0^{1,p}(\Omega) \cap W^{2,p}(\Omega))^3$ , for  $p > 3$ . This is also the approach of Marsden & Hughes [1978, p. 208], Valent [1978].

*Remark 2.4.* It is precisely the lack of available regularity result (for the linear elasticity system) in the case of a cylindrical domain such as  $\Omega^c$  and of mixed boundary conditions of the form (2.3)-(2.6) which limits

the applicability of the method to pure Dirichlet problems and domains with smooth boundaries.

*Remark 2.5.* Under the same assumptions, one can also prove the 1-1 character of the mapping

$$\varphi : x \in \Omega \rightarrow \varphi(x) = x + u(x),$$

a highly desirable property of the solution.

Regarding existence theory for nonlinear elasticity models, we mention the fundamental results of Ball [1977]. For yet another interesting approach, see Oden [1979].

We notice at this point that a variational formulation of equations (2.1)-(2.6), (2.17) consists in expressing that the pair  $(u, \sigma)$ , with  $u = (u_i)$  and  $\sigma = (\sigma_{ij})$ , satisfies :

$$(2.18) \quad u \in V^\epsilon \text{ def } \{v = (v_i) \in (W^1, P(\Omega^\epsilon))^3 ; v_1, v_2 \text{ are independent of } x_3 \\ \text{on } \Gamma_0^\epsilon, v_3 = 0 \text{ on } \Gamma_0^\epsilon\},$$

$$(2.19) \quad \sigma \in \Sigma^\epsilon \text{ def } \{\tau = (\tau_{ij}) \in (L^2(\Omega))^9 ; \tau_{ij} = \tau_{ji}\},$$

$$(2.20) \quad \forall v \in V^\epsilon, \int_{\Omega^\epsilon} \sigma_{ij} \gamma_{ij}(v) + \int_{\Omega^\epsilon} \sigma_{ij} \partial_i u_\ell \partial_j v_\ell = \\ = \int_{\Omega^\epsilon} f_3^\epsilon v_3 + \int_{\Gamma_+^\epsilon \cup \Gamma_-^\epsilon} g_3^\epsilon v_3 + \int_Y \left\{ \int_\epsilon^\epsilon v_\alpha dx_3 \right\} h_\alpha^\epsilon,$$

$$(2.21) \quad \forall \tau \in \Sigma^\epsilon, \epsilon^3 \int_{\Omega^\epsilon} (\Lambda \sigma)_{ij} \tau_{ij} - \int_{\Omega^\epsilon} \tau_{ij} \gamma_{ij}(u) - \frac{1}{2} \int_{\Omega^\epsilon} \tau_{ij} \partial_i u_\ell \partial_j u_\ell \\ - \epsilon^3 \sum_{2 \leq q \leq Q} a_{ijk_1 k_2 \dots k_{2q-1} k_{2q}} \int_{\Omega^\epsilon} \bar{\gamma}_{k_1 k_2}(u) \dots \bar{\gamma}_{2q-1, 2q}(u) \tau_{ij} = 0,$$

provided the number  $p$  is chosen to be large enough, so that all the integrals make sense.

*Remark 2.6.* Specific regularity assumptions on the data  $f_3^\epsilon$ ,  $g_3^\epsilon$ ,  $h_a^\epsilon$  will be made later on. For the time being, it suffices to assume that they are smooth enough so that all integrals appearing in (2.20)-(2.21) make sense.

3. DEFINITION OF A "LIMIT" PROBLEM FOR  $\epsilon = 0$ . Our first task is to define a problem equivalent to the variational problem (2.20)-(2.21), but now posed over a domain which does not depend on  $\epsilon$ . Accordingly, we shall successively define appropriate changes of variables and changes of functions. We let

$$\Omega = \omega \times [-1,1], \quad \Gamma_0 = \gamma \times [-1,1],$$

$$\Gamma_+ = \omega \times \{1\}, \quad \Gamma_- = \omega \times \{-1\},$$

and with each point  $X \in \bar{\Omega}$ , we associate the point  $X^\epsilon \in \bar{\Omega}^\epsilon$  through the correspondence

$$X = (x_1, x_2, x_3) \in \bar{\Omega} \rightarrow X^\epsilon = (x_1, x_2, \epsilon x_3) \in \bar{\Omega}^\epsilon.$$

With the space  $V^\epsilon$ ,  $\Sigma^\epsilon$  of (2.18)-(2.19), we associate the spaces

$$(3.1) \quad V = \{v = (v_i) \in (W^1, P(\Omega))^3 ; v_1, v_2 \text{ are independent of } x_3 \\ \text{on } \Gamma_0^\epsilon, v_3 = 0 \text{ on } \Gamma_0^\epsilon\}.$$

$$(3.2) \quad \Sigma = \{\tau = (\tau_{ij}) \in (L^2(\Omega))^9 ; \tau_{ij} = \tau_{ji}\}.$$

With the functions  $(v_i) \in V^\epsilon$ ,  $(\tau_{ij}) \in \Sigma^\epsilon$ , we associate the functions  $(v_i^\epsilon) \in V$ ,  $(\tau_{ij}^\epsilon) \in \Sigma$  defined by

$$(3.3) \quad v_\alpha(X^\epsilon) = v_\alpha^\epsilon(X), \quad v_3(X^\epsilon) = \epsilon v_3^\epsilon(X),$$

$$(3.4) \quad \tau_{\alpha\beta}(X^\epsilon) = \epsilon^{-1} \tau_{\alpha\beta}^\epsilon(X), \quad \tau_{\alpha 3}(X^\epsilon) = \tau_{\alpha 3}^\epsilon(X), \quad \tau_{33}(X^\epsilon) = \tau_{33}^\epsilon(X),$$

for all corresponding points  $X^\epsilon \in \bar{\Omega}^\epsilon$  and  $X \in \bar{\Omega}$ .

As regards the data, we shall assume that there exist functions  $f_3$ ,  $g_3$ ,  $h_\alpha$  which are independent of  $\epsilon$  such that

$$(3.5) \quad f_3^\epsilon(X^\epsilon) = f_3(X),$$

$$(3.6) \quad g_3^\epsilon(X^\epsilon) = \epsilon g_3(X),$$

$$(3.7) \quad h_\alpha^\epsilon(y) = \epsilon^{-1} h_\alpha(y) \text{ for all points } y \in \gamma.$$

The above relations have the basic effects that some integrals appearing in the variational formulation (2.20)-(2.21) of the three-dimensional problem are left unaltered, up to an appropriate multiplicative power of  $\epsilon$ . More specifically, one has

$$\int_{\Omega^\epsilon} \sigma_{ij}^\epsilon \gamma_{ij}(v) = \epsilon^2 \int_{\Omega} \sigma_{ij}^\epsilon \gamma_{ij}(v^\epsilon),$$

for all corresponding pairs  $(v, \sigma) \in V^\epsilon \times \Sigma^\epsilon$  and  $(v^\epsilon, \sigma^\epsilon) \in V \times \Sigma$ , and

$$\begin{aligned} \int_{\Omega^\epsilon} f_3^\epsilon v_3 + \int_{\Gamma_+^\epsilon \cup \Gamma_-^\epsilon} g_3^\epsilon v_3 + \int_Y \left\{ \int_{-\epsilon}^\epsilon v_\alpha dx_3 \right\} h_\alpha^\epsilon &= \\ &= \epsilon^2 \left\{ \int_{\Omega} f_3 v_3 + \int_{\Gamma_+ \cup \Gamma_-} g_3 v_3 + \int_Y \left\{ \int_{-1}^1 v_\alpha^\epsilon dt \right\} h_\alpha \right\}, \end{aligned}$$

for all corresponding functions  $v \in V^\epsilon$  and  $v^\epsilon \in V$ . Notice that the integrals appearing in the above equations precisely represent the classical duality (in elasticity theory) between the stresses and strains on the one hand, and between the forces and displacements on the other.

The justification of the scaling factor  $\epsilon^2$  is twofold : first, we want the asymptotic expansion (3.17) below to start with a factor of  $\epsilon^0$  and secondly we want equations (3.18)-(3.19) below to contain all the terms appearing in the equations found by the same process in the linear case (cf. Ciarlet & Destuynder [1979a]).

It is then a purely computational matter to establish the following result, whose interest is to formulate the three-dimensional plate problem in a form where the dependence on the parameter  $\epsilon$  is very simple :

**THEOREM 3.1.** Let  $(u^\epsilon, \sigma^\epsilon) \in V \times \Sigma$  be constructed from a solution  $(u, \sigma) \in V^\epsilon \times \Sigma^\epsilon$  of (2.20)-(2.21) through formulas (3.3)-(3.4). Then  $(u^\epsilon, \sigma^\epsilon)$  is solution of

$$(3.8) \quad \forall v \in V, \mathcal{B}(\sigma^\epsilon, v) + 2\mathcal{C}_0(\sigma^\epsilon, u^\epsilon, v) + 2\epsilon^2 \mathcal{C}_2(\sigma^\epsilon, u^\epsilon, v) = \mathcal{F}(v),$$

$$(3.9) \quad \forall \tau \in \Sigma, \mathcal{A}_0(\sigma^\epsilon, \tau) + \epsilon^2 \mathcal{A}_2(\sigma^\epsilon, \tau) + \epsilon^4 \mathcal{A}_4(\sigma^\epsilon, \tau) + \\ + \mathcal{B}(\tau, u^\epsilon) + \mathcal{C}_0(\tau, u^\epsilon, u^\epsilon) + \epsilon^2 \mathcal{C}_2(\tau, u^\epsilon, u^\epsilon) + \\ + \epsilon \mathcal{R}(\epsilon; \tau, u^\epsilon) = 0,$$

where, for arbitrary elements  $u, v \in V$  and  $\sigma, \tau \in \Sigma$ ,

$$(3.10) \quad \mathcal{B}(\tau, v) = - \int_{\Omega} \tau_{ij} \gamma_{ij}(v),$$

$$(3.11) \quad \mathcal{C}_0(\tau, u, v) = - \frac{1}{2} \int_{\Omega} \tau_{ij} \partial_i u_3 \partial_j v_3,$$

$$(3.12) \quad \mathcal{C}_2(\tau, u, v) = - \frac{1}{2} \int_{\Omega} \tau_{ij} \partial_i u_\alpha \partial_j v_\alpha,$$

$$(3.13) \quad \mathcal{F}(v) = - \left[ \int_{\Omega} f_3 v_3 + \int_{\Gamma_+ \cup \Gamma_-} g_3 v_3 + \int_Y \left\{ \int_{-1}^1 v_\alpha dt \right\} h_\alpha \right],$$

where the functions  $f_3, g_3, h_\alpha$  are those appearing in formulas (3.5)-(3.7),

$$(3.14) \quad \mathcal{A}_0(\sigma, \tau) = \int_{\Omega} \left[ \left( \frac{1+v}{E} \right) \sigma_{\alpha\beta} - \frac{v}{E} \sigma_{\mu\mu} \delta_{\alpha\beta} \right] \tau_{\alpha\beta},$$

$$(3.15) \quad \mathcal{A}_2(\sigma, \tau) = \int_{\Omega} \left\{ 2 \left( \frac{1+v}{E} \right) \sigma_{\alpha 3} \tau_{\alpha 3} - \frac{v}{E} (\sigma_{33} \tau_{\mu\mu} + \sigma_{\mu\mu} \tau_{33}) \right\},$$

$$(3.16) \quad \mathcal{A}_4(\sigma, \tau) = \frac{1}{E} \int_{\Omega} \sigma_{33} \tau_{33},$$

and  $\mathcal{R}(\epsilon, \tau, u^\epsilon)$  is a polynomial with respect to  $\epsilon$ , whose coefficients, which are integrals over  $\Omega$ , are independent of  $\epsilon$ . ■

Since the forms  $\mathcal{B}, \mathcal{C}_0, \mathcal{C}_2, \mathcal{F}, \mathcal{A}_0, \mathcal{A}_2, \mathcal{A}_4$ , are all independent of  $\epsilon$ , as well as the coefficients of the nonnegative powers of  $\epsilon$  in the polynomial  $\mathcal{R}(\epsilon, \tau, u^\epsilon)$ , and since  $\epsilon$  is thought of as being a "small" parameter, we are naturally led to define a formal series of "approximations" of a solution  $(u^\epsilon, \sigma^\epsilon)$  of (3.8)-(3.9) by letting *a priori* (the leading term  $(u, \sigma)$  in the following expansion should not be confused with a solution of the original three-dimensional problem) :

$$(3.17) \quad (u^\epsilon, \sigma^\epsilon) = (u, \sigma) + \epsilon(u^1, \sigma^1) + \epsilon^2(u^2, \sigma^2) + \dots$$

Then, following the principle of the asymptotic expansion method, we equate to zero the factors of the successive powers  $\epsilon^p$ ,  $p \geq 0$ , in the expressions obtained when the expansion (3.17) is used in (3.8)-(3.9).

In this fashion, we find :

- (i) equations to be satisfied by the first term ;
- (ii) recurrence relations for the following terms (of course, nothing guarantees at this stage the existence of the terms  $(u, \sigma)$ ,  $(u^1, \sigma^1)$ , etc., let alone the possible convergence of the series (3.17)).

In the sequel, we shall be concerned with the computation of the first term  $(u, \sigma)$  which, according to the above considerations, should satisfy :

$$(3.18) \quad \forall v \in V, \mathcal{B}(\sigma, v) + 2 \mathcal{C}_0(\sigma, u, v) = \mathcal{F}(v),$$

$$(3.19) \quad \forall \tau \in \mathbb{I}, \mathcal{A}_0(\sigma, \tau) + \mathcal{B}(\tau, u) + \mathcal{C}_0(\tau, u, u) = 0.$$

In this respect, our main results consist in :

- (i) establishing the existence of (at least) one solution to the "limit" problem (3.18)-(3.19) ;
- (ii) recognizing a known two-dimensional plate model in this same limit problem.

4. EQUIVALENCE OF THE "LIMIT" PROBLEM WITH THE VON KÁRMÁN EQUATIONS. We let

$$g_{3\pm}^{\epsilon} = g_3^{\epsilon} \text{ on } \Gamma_{\pm},$$

and we denote by  $\partial_{\nu}$  the exterior normal derivative operator along the boundary  $\gamma$  of the middle surface  $\omega$ . We first establish the equivalence of the "limit" problem (3.18)-(3.19) with a two-dimensional, displacement, model :

THEOREM 4.1. Assume that the data have the following regularity :

$$(4.1) \quad f_3 \in L^2(\Omega), \quad g_3 \in L^2(\Gamma_+ \cup \Gamma_-), \quad h_a \in H^{\frac{3}{2}}(\gamma),$$

and that the functions  $h_a$  verify the following compatibility conditions :

$$(4.2) \quad \int_{\gamma} h_1 = \int_{\gamma} h_2 = \int_{\gamma} (x_1 h_2 - x_2 h_1) = 0.$$

Eqs. (3.18)-(3.19) have at least one solution  $(u, \sigma) = ((u_i), (\sigma_{ij}))$  in the space  $V \times \Sigma$ , which is obtained as follows :

First, one solves the two-dimensional problem : Find  $u^0 = (u_i^0) : \omega \rightarrow \mathbb{R}^3$  such that

$$(4.3) \quad \frac{2E}{3(1-v^2)} \Delta^2 u_3^0 - 2\sigma_{\alpha\beta}^0(u^0) \partial_{\alpha\beta} u_3^0 = \left( g_{3+} + g_{3-} + \int_{-1}^1 f_3 dt \right) \text{ in } \omega,$$

$$(4.4) \quad \partial_{\alpha} \sigma_{\alpha\beta}^0(u^0) = 0 \text{ in } \omega,$$

$$(4.5) \quad u_3^0 = \partial_{\nu} u_3^0 = 0 \text{ on } \gamma,$$

$$(4.6) \quad \sigma_{\alpha\beta}^0(u^0) v_{\alpha} = h_{\beta} \text{ on } \gamma,$$

where

$$(4.7) \quad \begin{aligned} \sigma_{\alpha\beta}^0(u^0) &\stackrel{\text{def}}{=} \frac{E}{(1-v^2)} \{ (1-v) \gamma_{\alpha\beta}(u^0) + v \gamma_{\mu\mu}(u^0) \delta_{\alpha\beta} \} \\ &+ \frac{E}{2(1-v^2)} \{ (1-v) \partial_{\alpha} u_3^0 \partial_{\beta} u_3^0 + v \partial_{\mu} u_3^0 \partial_{\nu} u_3^0 \delta_{\alpha\beta} \}. \end{aligned}$$

This problem has at least one solution  $u^0 = (u_a^0, u_3^0)$  in the space  $(H^3(\omega))^2 \times (H_0^2(\omega) \cap H^4(\omega))$ .

Secondly, one defines, for  $(x_1, x_2, x_3) \in \Omega$ ,

$$(4.8) \quad u_3(x_1, x_2, x_3) = u_3^0(x_1, x_2),$$

$$(4.9) \quad u_\alpha = u_\alpha^0 - x_3 \partial_\alpha u_3^0,$$

$$(4.10) \quad \sigma_{\alpha\beta} = \sigma_{\alpha\beta}^0(u^0) - \frac{Ex_3}{(1-v^2)} \{ (1-v) \partial_{\alpha\beta} u_3^0 + v \Delta u_3^0 \delta_{\alpha\beta} \},$$

$$(4.11) \quad \sigma_{3\beta} = \sigma_{\beta 3} = - \frac{E(1-x_3^2)}{2(1-v^2)} \partial_\beta \Delta u_3^0,$$

$$(4.12) \quad \begin{aligned} \sigma_{33} &= \frac{(x_3+1)}{2} g_{3+} + \frac{(x_3-1)}{2} g_{3-} + \left\{ \frac{(1+x_3)}{2} \int_{-1}^1 f_3 dt - \int_{-1}^{x_3} f_3 dt \right\} \\ &\quad + \frac{Ex_3(1-x_3^2)}{6(1-v^2)} \Delta^2 u_3^0 - \frac{E(1-x_3^2)}{2(1-v^2)} \{ (1-v) \partial_{\alpha\beta} u_3^0 \partial_{\alpha\beta} u_3^0 + v (\Delta u_3^0)^2 \}. \end{aligned}$$

Conversely, any sufficiently regular solution of (3.18), (3.19) is necessarily of the form (4.8)-(4.12) with  $u^0 = (u_i^0)$  solution of problem (4.3)-(4.6). ■

Let us briefly sketch the main steps in the proof of this theorem.

Step 1 : In (3.19), we successively choose "trial" functions  $\tau \in \Gamma$  of the particular forms

$$(4.13) \quad \tau = (\tau_{ij}), \text{ with } \tau_{\alpha\beta} = \tau_{33} = 0,$$

$$(4.14) \quad \tau = (\tau_{ij}), \text{ with } \tau_{\alpha j} = 0.$$

Then if we restrict to solutions for which  $u_3$  is sufficiently regular, we find that, with the particular choices (4.13) and (4.14), (3.19) is satisfied if and only if :

$$(4.15) \quad \begin{cases} \text{the function } u_3 \text{ is independent of the variable } x_3 \text{ and it} \\ \text{can be identified with a function } u_3^0 \in H_0^2(\omega), \end{cases}$$

$$(4.16) \quad \exists u_a^0 \in H^1(\omega), \quad u_\alpha = u_\alpha^0 - x_3 \partial_\alpha u_3^0.$$

*Step 2 : Computation of the functions  $u_\alpha^0$  and  $u_3^0$  : In (3.18), (3.19), we successively choose "trial" functions of the particular forms*

$$(4.17) \quad \begin{cases} \tau_{\alpha\beta} = \tau_{\alpha\beta}^0 \in L^2(\omega), & \tau_{i3} = 0, \\ v_\alpha = v_\alpha^0 \in H^1(\omega), & v_3 = 0, \end{cases}$$

$$(4.18) \quad \begin{cases} \tau_{\alpha\beta} = x_3 \tau_{\alpha\beta}^1 & \tau_{\alpha\beta}^1 \in L^2(\omega), \\ v_\alpha = x_3 \partial_\alpha v, & v_3 = v, \quad v \in H_0^2(\omega). \end{cases}$$

After some computations (and elimination of the other unknowns) we find that the functions  $u_\alpha^0 \in H^1(\omega)$  and  $u_3^0 \in H_0^2(\omega)$  should be solution of (4.3)-(4.6).

*Step 3 : Computation of the stresses  $\sigma_{ij}$  : Once the functions  $\sigma_{\alpha\beta}^0(u^0)$  and  $u_3^0$  have been computed by solving (4.3)-(4.6), it turns out that (3.19) with  $\tau = (\tau_{ij})$ ,  $\tau_{i3} = 0$ , and (3.18) are satisfied if and only if the stresses  $\sigma_{ij}$  are given by (4.10)-(4.12).*

*Step 4 : Existence of a solution to the two-dimensional problem (4.3)-(4.6), for data possessing the regularity (4.1) : One can proceed in two ways :*

(i) The variational formulation of (4.3)-(4.6) amounts to finding the stationary points of the functional (we let  $v^0 = (v_i^0)$ )

$$(4.19) \quad \begin{aligned} J(v^0) = & \frac{E}{(1-v^2)} \int_{\omega} \left\{ \frac{1}{3} (\Delta v_3^0)^2 + (1-v) \gamma_{\alpha\beta} (v^0) \partial_\alpha v_3^0 \partial_\beta v_3^0 \right. \\ & + v \gamma_{\lambda\lambda} (v^0) \partial_\mu v_3^0 \partial_\mu v_3^0 + \frac{1}{4} (\partial_\alpha v_3^0 \partial_\alpha v_3^0)^2 \\ & \left. + (1-v) \gamma_{\alpha\beta} (v^0) \gamma_{\alpha\beta} (v^0) + v \gamma_{\lambda\lambda} (v^0) \gamma_{\mu\mu} (v^0) \right\} \\ & - \int_{\omega} (g_{3+} + g_{3-} + \int_{-1}^1 f_3 dt) v_3^0 - 2 \int_{\gamma} h_\alpha v_\alpha^0, \end{aligned}$$

when  $v^0$  varies over the space  $(H^1(\omega))^2 \times H_0^2(\omega)$ . Because of the compatibility conditions (4.2), this functional is also well-defined over the space

$$(4.20) \quad \mathcal{V} = \{(H^1(\omega))^2 / V^0\} \times H_0^2(\omega),$$

where

$$(4.21) \quad \begin{aligned} v^0 &= \{v = (v_\alpha) \in (H^1(\omega))^2 ; \gamma_{\alpha\beta}(v) = 0\} \\ &= \{v = (v_\alpha) \in (H^1(\omega))^2 ; \\ &\quad \exists a_\alpha, b \in \mathbb{R}, v_1 = a_1 - bx_2, v_2 = a_2 + bx_1\}, \end{aligned}$$

and besides, it is now coercive over this space (it is not coercive over the space  $(H^1(\omega))^2 \times H_0^2(\omega)$ ), i.e.,

$$\lim_{\|v^0\|_Y \rightarrow \infty} J(v^0) = +\infty,$$

provided the norms  $\|h_\alpha\|_{L^2(\omega)}$  are small enough.

In addition, it can be shown that the functional  $J$  is weakly lower semi-continuous over the space  $Y$ , and the conclusion follows by standard arguments.

(ii) In order to have an existence theory devoid of any restriction on the magnitude of the functions  $h_\alpha$ , one first introduces the so-called Airy stress function, as shown in Theorem 4.2 below (a process which again shows the necessity of imposing compatibility conditions on the functions  $h_1, h_2$ ). Next, one may use the existence theorem of John and Nečas [1975]. One can also eliminate the Airy stress function, following the method of Berger [1977], and show that the resulting problem in the single unknown  $u_3^0$  amounts to finding the stationary point of a specific functional, which has at least one minimum over the space  $H_0^2(\omega)$ , as in Rabier [1980].

Finally, using standard regularity results for the system of equations of linear, two-dimensional, elasticity, in conjunction with the method described in Lions [1969, p. 56], one can show that the solutions  $(u, \sigma)$  of problem (3.18), (3.19) found in the above process possess the following regularity :

$$u_3^0 \in H_0^2(\omega) \cap H^4(\omega), \quad u_\alpha^0 \in H_0^1(\omega) \cap H^3(\omega),$$

$$\sigma_{\alpha\beta} \in H^2(\Omega), \quad \sigma_{3\beta} \in H^1(\Omega), \quad \sigma_{33} \in L^2(\Omega).$$

*Remark 4.1.* Of course, it now remains to go back to the set  $\Omega^\epsilon$ , i.e., one must define functions on the set  $\Omega^\epsilon$ , which correspond to the functions  $u_i$  and  $\sigma_{ij}$  just constructed. For the sake of brevity, we shall skip the corresponding straightforward computations, simply based on formulas (3.3)-(3.7). It suffices to mention that their effect amounts to introducing appropriate powers of  $\epsilon$  at some places in the above equations. Thus for instance, equation (4.3), expressed with the "new" functions, now reads :

$$\frac{2E\epsilon^3}{3(1-\nu^2)} \Delta^2 u_3^0 = 2\epsilon \sigma_{\alpha\beta}^0 (u_3^0) \partial_{\alpha\beta} u_3^0 + (g_{3+}^\epsilon + g_{3-}^\epsilon + \int_{-\epsilon}^{\epsilon} f_3^\epsilon dx_3). \quad \blacksquare$$

An important conclusion to be drawn from the above theorem is that the expressions found for the functions  $u_i$  and  $\sigma_{ij}$  are identical to, or similar to, the assumed expressions found in the literature concerning nonlinear plate theory. In particular, we have obtained Kirchhoff-Love displacement fields, i.e., of the form (4.8), (4.9), whereas they are usually derived from an a priori assumption of a geometrical nature (cf. e.g. Washizu [1975, Eq. (8.60)]).

In the same fashion, the expressions found in (4.7) for the stresses  $\sigma_{\alpha\beta}^0$  (i.e.,  $\sigma_{\alpha\beta}$  for  $x_3 = 0$ ) are standard in nonlinear plate theory, where they are usually derived after a priori assumptions have been made regarding which terms should be neglected in the strain tensor corresponding to the two-dimensional problem (cf. e.g. Stoker [1968, pp. 42-47]). Likewise, the expressions found in (4.11) for the stresses  $\sigma_{3\beta}$  are similar to those found in Green and Zerna [1968, Eq. (7.7.3)], where they are assumed to be quadratic in  $x_3$ , etc.

In the second, and final, stage of our analysis, we establish the equivalence of problem (4.3)-(4.6) with the von Kármán equations (4.23)-(4.27). This equivalence essentially relies upon the introduction of the so-called Airy stress function  $\psi$ , which satisfies (4.22). We recall that the space  $V^0$  has been defined in (4.21).

In the next theorem, we assume that the set  $\omega$  is simply connected, and is of Nikodym type, in the sense of Deny and Lions [1953-1954] ; for instance, this is the case if the set  $\omega$  is star-shaped.

Without loss of generality, we also assume that the origin 0 belongs to the boundary  $\gamma$  of the set  $\omega$ . Given a point  $y$  along the boundary  $\gamma$ , we denote by  $\gamma(y)$  the arc joining the point 0 to the point  $y$  along  $\gamma$ .

**THEOREM 4.2.** Assume the data satisfy the regularity assumptions (4.1) and the compatibility conditions (4.2) and let there be given any solution

$$\mathbf{u}^0 = \left( (u_\alpha^0), u_3^0 \right) \in [H^3(\omega)]^2 \times [H_0^2(\omega) \cap H^4(\omega)]$$

of problem (4.3)-(4.6). Then there exists a function  $\varphi \in H^1(\omega)$ , uniquely determined if we impose  $\varphi(0) = \partial_1 \varphi(0) = \partial_2 \varphi(0)$ , such that

$$\partial_{11}\varphi = \sigma_{22}^0(u^0), \quad \partial_{12}\varphi = -\sigma_{12}^0(u^0), \quad \partial_{22}\varphi = \sigma_{11}^0(u^0).$$

Besides, the pair  $(\varphi, u_3^0)$  is solution of the von Kármán equations :

$$\frac{2E}{3(1-\nu^2)} \Delta^2 u_3^0 = 2[\varphi, u_3^0] + (g_{3+} + g_{3-} + \int_{-1}^1 f_3 dt) \text{ in } \omega,$$

$$\Delta^2 \varphi = -\frac{E}{2}[u_3^0, u_3^0] \text{ in } \omega,$$

$$u_3^0 = \partial_v u_3^0 = 0 \text{ on } \gamma,$$

$$\varphi(y) = -y_1 \int_{\gamma(y)} h_2 + y_2 \int_{\gamma(y)} h_1$$

$$+ \int_{\gamma(y)} (x_1 h_2 - x_2 h_1), \quad y \in \gamma,$$

$$\partial_v \varphi(y) = -v_1(y) \int_{\gamma(y)} h_2 + v_2(y) \int_{\gamma(y)} h_1, \quad y \in \gamma,$$

where, for any smooth enough functions  $v$  and  $w$ ,

$$[v, w] = \partial_{11}v \partial_{22}w + \partial_{22}v \partial_{11}w - 2\partial_{12}v \partial_{12}w.$$

Conversely, let there be given any solution

$$(\varphi, u^0) \in H^4(\omega) \times (H_0^2(\omega) \cap H^4(\omega))$$

of problem (4.23)-(4.27). Then, if we define functions  $\sigma_{\alpha\beta}^0$  by letting

$$\sigma_{11}^0 = \partial_{22}\varphi, \quad \sigma_{12}^0 = \sigma_{21}^0 = -\partial_{12}\varphi, \quad \sigma_{22}^0 = \partial_{11}\varphi,$$

there exists a unique element  $(u_\alpha^0)$  in the space  $(H^3(\omega))^2/V^0$  such that

$$\begin{aligned} \sigma_{\alpha\beta}^0 &= \frac{E}{(1-\nu^2)} \{ (1-\nu) \gamma_{\alpha\beta}(u^0) + \nu \gamma_{\mu\mu}(u^0) \delta_{\alpha\beta} \} \\ &\quad + \frac{E}{2(1-\nu^2)} \{ (1-\nu) \partial_\alpha u_3^0 \partial_\beta u_3^0 + \nu \partial_\mu u_3^0 \partial_\mu u_3^0 \delta_{\alpha\beta} \}, \quad u^0 = ((u_\alpha^0), u_3^0), \end{aligned}$$

and besides, the element  $u^0$  is solution of problem (4.3)-(4.6). ■

*Remark 4.2.* A fairly complete mathematical analysis of the von Kármán equations, regarding notably existence theory, multiplicity of solutions, bifurcation theory, etc ... , is found in Ciarlet & Rabier [1980]. ■

## 5. CONCLUSIONS

(i) The main conclusion is of course that we have been able to mathematically justify the derivation of a nonlinear plate model from a well-accepted three-dimensional nonlinear elasticity model, associated with specific boundary conditions along the lateral surface of the plate.

(ii) Which boundary conditions along the lateral surface are appropriate for the three-dimensional problem is a question of importance since different boundary conditions yields fundamentally different two-dimensional problems (as expected, of course, but this does not seem to be always clear in the literature). In this respect, see notably Ciarlet & Destuynder [1979b], where the case of a "clamped" plate is considered.

(iii) In order that a "limit" problem exist, it has been found that the various data should simultaneously vary in an appropriate manner as  $\epsilon$  approaches zero, as expressed by relations (2.16) and (3.5)-(3.7). These are not the only possible ones, however. For example, the Lamé coefficients  $\lambda, \mu$  appearing in (2.8) can stay constant provided relations (3.5)-(3.7) are replaced by the following :

$$f_3^\epsilon(x^\epsilon) = \epsilon^3 f_3(x), \quad g_3^\epsilon(x^\epsilon) = \epsilon^3 g_3(x),$$

$$h_\alpha^\epsilon(y) = \epsilon^2 h_\alpha^\epsilon(y) \text{ for all } y \in \gamma,$$

and the "higher order constants" appearing in the constitutive equation decay sufficiently rapidly with  $\epsilon$ . Then it is readily verified that the same "limit problem" (3.18)-(3.19) is retained by an application of the asymptotic expansion method. The above relations are much less realistic however if body forces, such as the weight, are to be taken into account. But they cannot be disposed of : One cannot expect a plate of zero thickness to carry any load !

The interpretation of relations (2.16) is simple : They express that

the rigidity of the constitutive material of the plate should increase as the thickness of the plate approaches zero, if we are to find a "limit" model compatible with relations (3.5)-(3.7). Incidentally, similar conclusions have been reached in a related linear problem by Cailleurie [1980].

The assumption that the coefficients corresponding to the higher-order terms in the constitutive equation (2.17) are constant is in turn made necessary by the requirement that these terms do not appear in, and thus do not affect, the "limit" problem. A different limit problem would otherwise result which could be studied for its own sake. Our aim was however to clearly delineate three-dimensional constitutive equations correspond to *precisely* the von Kármán equations. Notice also that the above assumption regarding the "higher order constants" is evidently satisfied if the constitutive equation is linear (as a relation between the tensors  $\sigma$  and  $\bar{\gamma}(u)$ ), as was the case in Ciarlet [1980].

(iv) The present analysis suggests that we consider the von Kármán equations, together with the expressions simultaneously found for the unknowns  $u_i$ ,  $\sigma_{ij}$  as forming a *consistent set of approximations* to the original three-dimensional problem, in the sense that these equations and expressions are all obtained as the solution of a single, three-dimensional problem, namely problem (3.18), (3.19).

Equivalently, if we start out with a solution of either two-dimensional problem, we may think of the expressions giving the unknowns  $u_i$ ,  $\sigma_{ij}$  as being the natural *extension* of this solution into the space  $V \times \mathbb{L}$ . Such an extension may prove useful for obtaining existence results for the original three-dimensional problem.

ACKNOWLEDGEMENTS. The author expresses his thanks to Professors C. TRUESDELL and S.S. ANITMAN for suggesting the consideration of general constitutive equations. This paper was written while the author was visiting the Mathematics Department of Cornell University, Ithaca, N.Y. ; in this respect, the financial support of the Exchange Visitor Program No. P - 1 - 43 is gratefully acknowledged.

## REFERENCES

BALL, J.M. [1977], Convexity conditions and existence theorems in nonlinear elasticity, *Arch. Rational Mech. Anal.* 63, 337-403.

BERGER, M.S. [1977], *Nonlinearity and Functional Analysis*, Academic Press, New York.

CAILLERIE, D. [1980], The effect of a thin inclusion of high rigidity in an elastic body, *Mathematical Methods in the Applied Sciences* (to appear).

CIARLET, P.G. [1980], A justification of the von Kármán equations, *Arch. Rational Mech. Anal.* 73, 349-389.

CIARLET, P.G. ; DESTUYNDER, P. [1979a], A justification of the two-dimensional linear plate model, *J. Mécanique* 18, 315-344.

CIARLET, P.G. ; DESTUYNDER, P. [1979b], A justification of a nonlinear model in plate theory, *Comput. Methods Appl. Mech. Engrg.* 17/18, 227-258.

CIARLET, P.G. ; KESAVAN, S. [1980], Two-dimensional approximations of three-dimensional eigenvalue problems in plate theory, *Comput. Methods Appl. Mech. Engrg.* (to appear).

CIARLET, P.G. ; RABIER, P. [1980], *Les Equations de von Kármán*, Lecture Notes in Mathematics, Vol. 826, Springer-Verlag, Heidelberg.

DENY, J. ; LIONS, J.-L. [1953-1954], Les espaces du type de Beppo-Levi, *Ann. Institut Fourier (Grenoble)* 4, 305-370.

DESTUYNDER, P. [1979], *Sur une Justification Mathématique des Théories de Plaques et de Coques en Élasticité Linéaire*, Doctoral Dissertation, Université Pierre et Marie Curie, Paris.

GREEN, A.E. ; ZERNA, W. [1968], *Theoretical Elasticity*, University Press, Oxford.

JOHN, F. [1971], Refined interior equations for thin elastic shells, *Comm. Pure Applied Math.* XXIV, 583-615.

JOHN, O. ; NEČAS, J. [1975], On the solvability of von Kármán equations, *Appl. Mat.* 20, 48-62.

LIONS, J.-L. [1969], *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris.

LIONS, J.-L. [1973], *Perturbations Singulières dans les Problèmes aux Limites et en Contrôle Optimal*, Lecture Notes in Mathematics, Vol. 323, Springer, Berlin.

MARSDEN, J.E. ; HUGHES, T.J.R. [1978], Topics in the mathematical foundations of elasticity, in *Nonlinear Analysis and Mechanics, Heriot-Watt Symposium, Volume II*, Pitman, London.

MURNAGHAN, F.D. [1937], Finite deformations of an elastic solid, *American Journal of Mathematics* 59, 235-260.

NOVOZHILOV, V.V. [1953], Foundations of the Nonlinear Theory of Elasticity, Graylock Press, Rochester.

ODEN, J.T. [1979], Existence theorems for a class of problems in nonlinear elasticity, *J. Math. Anal. Appl.* 69, 51-83.

RABIER, P. [1980], Doctoral Dissertation, Université Pierre et Marie Curie, Paris .

RIGOLOT, A. [1977], *Déplacements Finis et Petites Déformations des Poutres Droites : Analyse Asymptotique de la Solution à Grande Distance des Bases*, *J. Mécanique Appliquée* 1, 175-206.

STOKER, J.J. [1968], *Nonlinear Elasticity*, Gordon and Breach, New York.

TRUESDELL, C. ; NOLL, W. [1965], *The Non-Linear Field Theories of Mechanics*, Handbuch der Physik, Vol. III/3, Springer, Berlin.

VALENT, T. [1978], Teoremi di esistenza e unicità in elastostatica finita, *Rend. Sem. Mat. Univ. Padova* 60, 165-181.

VOIGT, W. [1893-1894], Ueber eine anscheinend notwendige Erweiterung der Theorie der Elastizität, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, pp. 534-552 (1893) ; pp. 33-42 (1894).

WANG, C.-C. ; TRUESDELL, C. [1973], *Introduction to Rational Elasticity*, Noordhoff, Groningen, 1973.

WASHIZU, K. [1975], *Variational Methods in Elasticity and Plasticity*, Second Edition, Pergamon, Oxford.

## CHANGING MESHES IN TIME-DEPENDENT PROBLEMS

Todd Dupont, University of Chicago

This note gives a brief summary of some results on finite element methods for evolution equations that use functions spaces that change with time. Most of these results are given in detail in [1].

In several areas of science and engineering, time-dependent problems arise which have solutions that are near-shocks in the sense that the solutions are smooth over most of the region but almost discontinuous in a small part of the region. If, as is frequently the case, the small area of roughness sweeps out a significant portion of the region during the life of the problem, then an approximate solution can be quite expensive to compute. The expense comes from the fact that a fine grid is needed in the region of roughness, and with a fixed grid that implies a fine grid over a large part of the region.

An example of such a problem is

$$(1) \quad \begin{aligned} u_t + v \cdot \nabla u - \nabla \cdot D\nabla u &= 0 && \text{on } \Omega \times (0, T], \\ u(x, 0) &= u_0(x) && \text{on } \Omega, \\ D\nabla u \cdot v &= g && \text{on } \partial\Omega \times (0, T], \end{aligned}$$

where  $\Omega$  is a bounded domain in  $\mathbb{R}^d$  with a smooth boundary and  $v$  is the outward normal. The function  $D$  is assumed to be positive throughout  $\bar{\Omega}$ .

### A Galerkin Method

Suppose that  $\mathcal{M}(t)$  is a finite-dimensional subspace of  $H^1(\Omega)$  for each  $t$  in  $[0, T]$  and that  $\mathcal{M}(t)$  varies smoothly except at a finite number of points  $T_j$ . One can define a Galerkin approximation to  $u$  to be a function  $U : [0, T] \rightarrow \bigcup \mathcal{M}(t)$ , where  $U(t) \in \mathcal{M}(t)$  and where  $U$  satisfies the usual Galerkin orthogonalities at each time:

$$(2) \quad \int_{\Omega} [(U_t + v \cdot \nabla U)\psi + \nabla U \cdot \nabla \psi] dx = \int_{\partial\Omega} g\psi d\sigma, \quad \psi \in \mathcal{M}(t).$$

At those points  $T_j$  at which  $\mathcal{M}$  changes discontinuously, use the  $L^2(\Omega)$ -projection into  $\mathcal{M}(T_j)$  of the limit from below to get  $U(T_j)$  to re-start this process.

### Quasi-Optimality

Using the above-defined process, started from the  $L^2(\Omega)$ -projection of  $u_0$ , one gets a quasi-optimality result

$$(3) \quad ||| U - u ||| \leq C \inf ||| \psi - u |||,$$

where the inf is taken over all functions  $\psi(t) \in \mathcal{M}(t)$  that vary smoothly except at the points  $T_j$  and are such that  $\psi(T_j)$  is the  $L^2(\Omega)$ -projection into  $\mathcal{M}(T_j)$  of the limit from below. The norm  $||| \cdot |||$  in (3) is one that is naturally associated with energy estimates for problems of the form of (1); it involves the maximum in time of the  $L^2(\Omega)$ -norm, the  $L^2(\Omega \times (0, T))$ -norm of the spatial gradient, and a semi-norm induced by the spaces  $\mathcal{M}(t)$ .

Estimates of the form of (3) are done in [1] for discrete-time processes, with the addition of a time discretization term to the right-hand side.

### **Nonconvergence**

If the mesh changes in a completely uncontrolled way the solution of (2) can converge to the wrong function as the meshes are made finer and finer. Each mesh change corresponds to adding a very small amount of dissipation, and thus if the mesh changes extremely often the solution will be smeared out.

### **Moving Finite Elements**

Take  $\Omega = (0,1)$  and suppose that  $\mathcal{M}(t)$  consists of the space of all continuous piecewise polynomials of degree  $\leq r$  over a mesh

$$0 = s_0 < s_1(t) < \dots < s_{N-1}(t) < s_N = 1.$$

The MFE method of K. Miller and R. Miller then uses the orthogonalities in (2) plus a rule that is formally derived by saying that the time-derivatives of the points  $s_j(t)$  are taken so as to minimize the  $L^2(\Omega)$ -norm of the residual plus a penalty term. (Such a calculation is purely motivational since the residual is in general not in  $L^2(\Omega)$ .) The penalty term is used to get nonsingular evolution equations and to control the spacing of the points  $s_j$ .

In the MFE process the grid points move with the solution and cluster around areas of roughness, thereby significantly decreasing the work to approximate near-shock solutions when compared to a fixed mesh method. In [2] there is a collection of interesting examples of the application of this method.

In [1] it is shown that, under appropriate hypotheses, the MFE behaves at least as well as a fixed grid process. This is clearly just a first step and does not explain the experimental success of the procedure. More recently the results of [1] have been extended to some multi-dimensional problems.

**References**

- [1] T. Dupont, Mesh modification for evolution equations, submitted to Math. Comp.
- [2] R.J. Gelinas, S.K. Doss, and K. Miller, The moving finite element method: applications to general partial differential equations with multiple large gradients, to appear J. Comp. Phys.
- [3] K. Miller and R. Miller, Moving finite elements, part I, to appear in SIAM J. Numer. Anal.

ALTERNATING-DIRECTION GALERKIN METHODS  
FOR PARABOLIC, HYPERBOLIC AND  
SOBOLEV PARTIAL DIFFERENTIAL EQUATIONS

Richard E. Ewing  
Mobil Research and Development Corporation  
Field Research Laboratory  
P. O. Box 900  
Dallas, Texas 75221

Abstract

A survey of some recent results in the use of alternating-direction finite element methods for linear and nonlinear partial differential equations of parabolic, hyperbolic, and Sobolev type is presented. These equations have applications to fluid flow in porous media, thermodynamics, wave propagation, nonlinear viscoelasticity, and hydrodynamics. The use of alternating-direction or operator-splitting methods will reduce multidimensional problems to repeated solution of one-dimensional problems. Thus optimal order work estimates can be obtained in all cases. Other new high-order and computationally efficient time-stepping procedures are also discussed and used as base schemes for the alternating-direction variants.

ALTERNATING-DIRECTION GALERKIN METHODS FOR PARABOLIC,  
HYPERBOLIC, AND SOBOLEV PARTIAL DIFFERENTIAL EQUATIONS

1. INTRODUCTION

In this paper, we shall present a survey of some recent results in the use of alternating-direction Galerkin methods for a variety of partial differential equations. We shall discuss methods for time-stepping partial differential equations of parabolic, hyperbolic, and Sobolev types in two and three spatial dimensions. The use of alternating-direction or operator-splitting methods will reduce multidimensional problems to repeated solution of one-dimensional problems. Thus optimal order work estimates can be obtained in all alternating-direction methods.

We shall basically consider only Galerkin or finite element alternating-direction (henceforth called AD) methods in this paper. Similar results can also be obtained for finite difference versions of our methods. Since the analysis of our methods will appear elsewhere, we shall only describe the methods in this manuscript and reference the analysis.

Alternating-direction methods were first used for time-dependent problems in the context of reservoir engineering models for fluid flow in porous media. The methods were developed in order to treat large scale multidimensional problems in a one-dimensional fashion on the small early-generation computers. Finite difference methods were developed for linear parabolic problems and analyzed thoroughly by Douglas, Peaceman, Rachford and others (see [10, 17, 18, 32]). Later Douglas and Dupont developed and analyzed a Laplace-modified Galerkin AD method for parabolic and hyperbolic equations with certain nonlinearities in [12]. These ideas were extended to stronger nonlinearities by Dendy in [8] and to unions of rectangular regions by Dendy and Fairweather in [9]. Then in [26, 27] Hayes extended these results to non-rectangular regions via patch approximations. In [28] Hayes and Percell extended these results to nonlinear capacity terms. Finally, in [11], Douglas discussed the combination of the results of [12, 28] with some of the iterative stabilization techniques presented in [14] to obtain other effective AD time-stepping procedures.

In this paper we shall discuss some recent advances in several different directions. First we discuss a tensor product projection of the solution into our computational subspaces and approximation theory results which greatly relax the smoothness assumptions required for all the earlier analysis of AD methods. Then we discuss some higher-order multistep time-stepping procedures which yield second, third, and in special cases fourth order time-truncation errors for parabolic problems. Previously, only second order methods with fairly strenuous coefficient constraints were known. We then extend the AD ideas to various partial differential equations of Sobolev type which are used in fluid flow in fractured media, thermodynamics, vibrational problems, nonlinear viscoelasticity, and hydrodynamics (see [6, 7, 25, 29, 30, 31, 33, 34]). Finally we present some direct methods and iterative stabilization techniques which yield new, high-order and computationally efficient methods.

Let  $\Omega$  be a bounded domain in  $\mathbb{R}^d$ ,  $2 \leq d \leq 3$ , with boundary  $\partial\Omega$ , and let  $J = (0, T)$ . We shall consider partial differential equations for  $u = u(x, t)$  of the form

$$(1.1) \quad \begin{aligned} a) \quad & e(x, u) \frac{\partial^2 u}{\partial t^2} + c(x, u) \frac{\partial u}{\partial t} - \nabla \cdot (a(x, u) \nabla u + b(x, u) \nabla \frac{\partial u}{\partial t} \\ & + g(x, u) \nabla \frac{\partial^2 u}{\partial t^2}) = f(x, t, u) \quad , \quad x \in \Omega, \quad t \in J, \\ b) \quad & u(x, t) = 0 \quad , \quad x \in \partial\Omega, \quad t \in J, \\ c) \quad & u(x, 0) = u_0(x) \quad , \quad x \in \Omega, \end{aligned}$$

for various choices of  $a$ ,  $b$ ,  $c$ ,  $e$ , and  $g$ . If  $e > 0$  and  $g > 0$ , we must also specify an additional initial condition of the form

$$(1.2) \quad u_t(x, 0) = v(x) \quad , \quad x \in \Omega.$$

If  $e = b = g = 0$  in (1.1) above, the equation is of parabolic type. This survey includes recent joint work by Jim Bramble and the author [3, 4] on problems of this type. If  $e > 0$  and  $c \leq b \leq g \leq 0$ , the problems are of hyperbolic type. If  $e > 0$  and either  $b > 0$  or  $g > 0$ , the problems are of

Sobolev type. Joint work with Linda Hayes [22, 23] on problems of this type will be discussed.

In Section 2 we shall present some preliminaries and notation. We then illustrate the basic ideas of AD methods for various cases with constant coefficients in Section 3. In Section 4 we shall discuss higher-order direct methods which use the ideas of [8, 12, 26, 27, 28]. In Section 5 we discuss iterative stabilization ideas which use the ideas of [13, 14, 19, 20, 24]. We also discuss certain computational aspects of these methods.

### III. PRELIMINARIES AND NOTATION

Let  $(u,v) = \int_{\Omega} uv dx$  and  $\|u\|^2 = (u,u)$ . Let the norm on the Sobolev space  $W^{k,p}(\Omega)$  be denoted by  $\|u\|_{k,p}$ , with the second index being suppressed if  $p = 2$ . Assume that  $\partial\Omega$  is Lipschitz continuous. Assume that the coefficients and solutions are smooth; we refer to the various papers referenced for more precisely defined constraints.

For  $h$  from a sequence of small positive numbers, let  $\{M_h^{\circ}[0,1]\}$  be a family of finite-dimensional subspaces of  $W^{1,\infty}(0,1)$  which vanish at  $x = 0$  and  $x = 1$  and which satisfy:

For some integer  $r > 2$  and some constant  $K_0$  and any

$$\begin{aligned} & \exists \epsilon W^{q,2}(0,1) \cap W^{1,\infty}(0,1), \\ & \inf_{x \in M_h^{\circ}[0,1]} [\|\phi - x\|_0 + h\|\phi' - x'\|_1 + h^{3/2} (\|\phi'' - x''\|_{0,\infty} + h\|\phi''' - x''' \|_{1,\infty})] \\ (2.1) \quad & \leq K_0 \|\phi\|_q h^q \end{aligned}$$

for  $1 \leq q \leq r + 1$ .

An example of a family of subspaces satisfying (2.1) is the continuous subspace of piecewise polynomials of degree at most  $r$  on each subinterval of length  $h$  of a uniform partition of  $[0,1]$ .

We next define one-dimensional projection operators  $P_x$ ,  $P_y$ , and  $P_z$ :

$W^{1,\infty} \rightarrow M_h^{\circ}[0,1]$  by

$$\begin{aligned} (2.2) \quad a) \quad & \int_0^1 \frac{\partial}{\partial x} (P_x u - u) \frac{\partial}{\partial x} x \, dx = 0 \quad , \quad x \in M_h^{\circ}[0,1], \\ b) \quad & \int_0^1 \frac{\partial}{\partial y} (P_y u - u) \frac{\partial}{\partial y} x \, dy = 0 \quad , \quad x \in M_h^{\circ}[0,1], \\ c) \quad & \int_0^1 \frac{\partial}{\partial z} (P_z u - u) \frac{\partial}{\partial z} x \, dz = 0 \quad , \quad x \in M_h^{\circ}[0,1]. \end{aligned}$$

Next, let  $I_d$  denote the unit cube in  $\mathbb{R}^d$  and define a sequence of subspaces on  $I_3$  by

$$(2.3) \quad M_h \equiv M_h [I_3] \equiv \overset{\circ}{M_h} [0,1] \times \overset{\circ}{M_h} [0,1] \times \overset{\circ}{M_h} [0,1].$$

We henceforth assume that  $\Omega = I_3$  (or  $I_2$  in  $\mathbb{R}^2$ ). See [9, 27] for techniques to extend these results to more general regions. We then define the three-dimensional tensor product projection  $Z = P_x P_y P_z u$  in  $M_h$ . Note that the one-dimensional operators commute and thus can be taken in any order. Using (1.1.b), we can then obtain a very important orthogonality result.

Lemma 2.1: If  $d = 2$  or  $d = 3$ , respectively,

$$(2.4) \quad \begin{aligned} a) \quad & \left( \frac{\partial^2}{\partial x \partial y} (P_x P_y u - u), \frac{\partial^2}{\partial x \partial y} \chi \right) = 0 \quad , \quad \chi \in M_h [I_2], \\ b) \quad & \left( \frac{\partial^3}{\partial x \partial y \partial z} (P_x P_y P_z u - u), \frac{\partial^3}{\partial x \partial y \partial z} \chi \right) = 0, \quad , \quad \chi \in M_h [I_3]. \end{aligned}$$

We next define some other projections into  $M_h$ . If  $a(x, u)$ ,  $b(x, u)$ , and  $g(x, u)$  are bounded below by positive constants, let  $w_a$ ,  $w_b$ , and  $w_g$  be the weighted elliptic projections satisfying:

$$(2.5) \quad \begin{aligned} a) \quad & (a(x, u) \nabla (w_a - u), \nabla \chi) = 0 \quad , \quad \chi \in M_h, \\ b) \quad & (b(x, u) \nabla (w_b - u), \nabla \chi) = 0 \quad , \quad \chi \in M_h, \\ c) \quad & (g(x, u) \nabla (w_g - u), \nabla \chi) = 0 \quad , \quad \chi \in M_h. \end{aligned}$$

Then, using the super-close approximation properties of the Galerkin solution in  $W^{1,2}$  and Lemma 3.1 of [16], we obtain the following important result:

Lemma 2.2: For  $Z = P_x P_y P_z u$  and  $w_a$ ,  $w_b$ , and  $w_g$  defined in (2.5), we have for some  $K_0 > 0$ ,

$$(2.6) \quad \|W_a - Z\|_1 + \|W_b - Z\|_1 + \|W_g - Z\|_1 \leq K_o \|u\|_{r+1} h^{r+1}.$$

Proof: (see [3]).

For  $k > 0$ , let  $N = T/k \in \mathbb{Z}$  and  $t^\sigma = \sigma k$ ,  $\sigma \in \mathbb{R}$ . Also let  $\phi^n \equiv \phi^n(x) \equiv \phi(x, t^n)$ . Define the following backward difference operators:

$$(2.7) \quad \begin{aligned} a) \quad \delta\phi^n &= \phi^n - \phi^{n-1} \\ b) \quad \delta^2\phi^n &= \phi^n - 2\phi^{n-1} + \phi^{n-2} \\ c) \quad \delta^3\phi^n &= \phi^n - 3\phi^{n-1} + 3\phi^{n-2} - \phi^{n-3} \\ d) \quad \delta^4\phi^n &= \phi^n - 4\phi^{n-1} + 6\phi^{n-2} - 4\phi^{n-3} + \phi^{n-4}. \end{aligned}$$

### III. DESCRIPTION OF THE METHODS - CONSTANT COEFFICIENTS

In this section we shall describe various methods for efficiently time-stepping the Galerkin spatial procedures for various forms of (1.1) with constant coefficients. We first consider the parabolic case of (1.1) where  $e \equiv b \equiv g \equiv 0$  and  $c$  and  $a$  are positive constants:

$$c \frac{\partial u}{\partial t} - a \Delta u = f(x, t, u).$$

For this case, we first present several multistep methods which will form our base schemes. Next, we shall introduce terms which allow us to use AD ideas in space.

For various special choices of parameters, we define the following class of backward differentiation, multistep, discrete time methods. Let

$U: \{t_0, \dots, t_N\} \rightarrow M_h$  be an approximate solution of (1.1). Assume that  $U^k$  are known for  $k \leq n$ . Given a desired global time-truncation error of order  $k^\mu$ ,  $\mu = 1, 2, 3, 4$ , we choose parameters  $\alpha_i(\mu)$ ,  $i = 1, 2, 3$ , and  $\beta(\mu)$  and an extrapolation operator  $E(\mu)$  for  $f(x, t, u)$  to define a method for determining  $U^{n+1}$  which satisfies

$$(3.1) \quad \begin{aligned} & k^{-1} (c \delta U^{n+1}, x) + \beta (a \nabla U^{n+1}, \nabla x) \\ & = k^{-1} (c [\alpha_1 \delta U^n + \alpha_2 \delta U^{n-1} + \alpha_3 \delta U^{n-2}], x) \\ & + \beta (f(t^{n+1}, E(\mu) U^{n+1}), x), \quad x \in M_h. \end{aligned}$$

Choices of the parameters and extrapolation operator for  $\mu = 1, \dots, 4$  are given in Table 1. By extrapolating the values of  $U^k$  in the nonlinear term  $f$ , we have produced a linear operator equation for  $U^{n+1}$  in terms of previous known values of  $U^k$ ,  $k \leq n$ . See [5, 21] for a detailed analysis of the stability and accuracy of these methods. We note that the case for  $\mu = 2$  is not the second-order Crank-Nicolson method which has a characteristic bounce. Instead, all the methods presented here are dissipative and strongly stable.

We next consider AD variants of (3.1). Let  $U^{n+1}$  satisfy

$$\begin{aligned}
 (3.2) \quad & k^{-1} (c\delta U^{n+1}, x) + \beta (a \nabla U^{n+1}, \nabla x) + \frac{k\beta^2 a^2}{c} \left[ \left( \frac{\partial^2}{\partial x \partial y} D(\mu) U^{n+1}, \frac{\partial^2}{\partial x \partial y} x \right) \right. \\
 & \left. + \left( \frac{\partial^2}{\partial x \partial z} D(\mu) U^{n+1}, \frac{\partial^2}{\partial x \partial z} x \right) + \left( \frac{\partial^2}{\partial y \partial z} D(\mu) U^{n+1}, \frac{\partial^2}{\partial y \partial z} x \right) \right] \\
 & + \frac{k^2 \beta^3 a^3}{c^2} \left( \frac{\partial^3}{\partial x \partial y \partial z} D(\mu) U^{n+1}, \frac{\partial^3}{\partial x \partial y \partial z} x \right) \\
 & = k^{-1} (c [\alpha_1 \delta U^n + \alpha_2 \delta U^{n-1} + \alpha_3 \delta U^{n-2}], x) \\
 & + \beta (f(t^{n+1}, E(\mu) U^{n+1}), x), \quad x \in M_h,
 \end{aligned}$$

where the operator  $D(\mu)U^{n+1}$  makes the additional terms "small" enough so as not to increase the order of the errors already present in the approximations. For example, for  $\mu = 1$  or  $\mu = 2$ , the choice  $D(\mu)U^{n+1} = \delta U^{n+1}$  will yield convergent schemes. For  $\mu = 3$ , we shall use  $D(3)U^{n+1} = \delta^2 U^{n+1}$ . For the case  $\mu = 4$ , the choice  $D(4)U^{n+1} = \delta^3 U^{n+1}$  would make the perturbation terms small enough for proper truncation error analysis, but will cause the method to be unstable. Instead, we shall choose

$$(3.3) \quad D(4) = \delta^2 U^{n+1} - c S_Y^{-1} \delta^2 U^n$$

with

$$\begin{aligned}
 (3.4) \quad S_Y &= \left( 1 + kY \left[ \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z} \right] \right. \\
 &\quad \left. + k^2 Y^2 \left[ \frac{\partial^2}{\partial x \partial y} + \frac{\partial^2}{\partial x \partial z} + \frac{\partial^2}{\partial y \partial z} \right] + k^3 Y^3 \frac{\partial^3}{\partial x \partial y \partial z} \right) \\
 &= \left( 1 + kY \frac{\partial}{\partial x} \right) \left( 1 + kY \frac{\partial}{\partial y} \right) \left( 1 + kY \frac{\partial}{\partial z} \right).
 \end{aligned}$$

Since  $c S_Y^{-1}$  is comparable to the identity operator, this choice of  $D(4)$  acts like  $\delta^3 U^{n+1}$ , and  $Y$  is chosen sufficiently large to make the method stable.

The additional terms in (3.2) allow the operator to factor in a manner exactly as in (3.4) into a sequence of one-dimensional operators. Since the methods presented in (3.2) involve up to five time levels, special start-up procedures must be discussed. Higher-order start-up procedures for the methods described in (3.1) have been presented and analyzed in [4]; however, the procedures have not been shown to be effective for AD methods. Start-up procedures for cases  $\mu = 1, 2, 3$  will appear in [3], but no procedure has been analyzed for the case  $\mu = 4$  at this time. The AD methods of (3.2) yield the same order convergence rates as the multistep methods of (3.1) but yield optimal order work estimates as well.

Next, we consider other partial differential equations by making different choices of coefficients in (1.1). If  $a > 0$ ,  $e > 0$ , and  $c \geq b \geq g \geq 0$ , we have an equation of hyperbolic type:

$$e \frac{\partial^2 u}{\partial t^2} - \nabla \cdot (a(x, u) \nabla u) = f(x, t, u).$$

AD methods of the form with  $d = 2$

$$(3.5) \quad \begin{aligned} & k^{-2} (e \delta^2 u^{n+1}, x) + (a \nabla u^n, \nabla x) + \lambda (\nabla \delta^2 u^{n+1}, \nabla x) \\ & + \frac{\lambda^2 k^2}{e} \left( \frac{\partial^2}{\partial x \partial y} \delta^2 u^{n+1}, \frac{\partial^2}{\partial x \partial y} x \right) = (f(t^n, u^n), x), x \in M_h, \end{aligned}$$

have been presented and analyzed in [8, 12]. The Laplace-modified ideas were presented and analyzed for both parabolic and hyperbolic equations in [12] and yield second order time-truncation estimates. Extensions to higher dimensions are straightforward as pointed out in [8]. However, since only the weighted elliptic projection (2.5.a) was used in the analysis, more smoothness on  $u$  was required than if  $Z = P_x P_y P_z u$  and Lemmas 2.1 and 2.2 had been used.

Next we discuss results for equations of Sobolev type which will appear in [23]. We first consider the case with  $a > 0$ ,  $b > 0$  and  $c > 0$  with  $e \geq g \geq 0$  in (1.1):

$$c \frac{\partial u}{\partial t} - \nabla \cdot (a \nabla u + b \nabla \frac{\partial u}{\partial t}) = f(x, t, u).$$

Equations of this form are studied in [19, 34]. Since equations of Sobolev type have a time derivative in the highest-order terms, they are in general inherently more stable than corresponding parabolic equations. However, the time derivatives in the highest-order terms also make the perturbation terms needed for AD variants much larger. Therefore three time levels will be required for  $O(k)$  accuracy and four levels for  $O(k^2)$  accuracy in this case. One method which has time-truncation errors of order  $k$  is:

$$\begin{aligned}
 & k^{-1} (c (U^{n+1} - U^{n-1}), x) + (a \nabla U^n, \nabla x) \\
 & + k^{-1} (b \nabla (U^{n+1} - U^{n-1}), \nabla x) \\
 (3.6) \quad & + \frac{b^2}{kc} \left[ \left( \frac{\partial^2}{\partial x \partial y} \delta^2 U^{n+1}, \frac{\partial^2}{\partial x \partial y} x \right) + \left( \frac{\partial^2}{\partial x \partial z} \delta^2 U^{n+1}, \frac{\partial^2}{\partial x \partial z} x \right) \right. \\
 & \left. + \left( \frac{\partial^2}{\partial y \partial z} \delta^2 U^{n+1}, \frac{\partial^2}{\partial y \partial z} x \right) \right] + \frac{b^3}{kc^2} \left( \frac{\partial^3}{\partial x \partial y \partial z} \delta^2 U^{n+1}, \frac{\partial^3}{\partial x \partial y \partial z} x \right) \\
 & = (f(t^n, U^n), x) \quad , x \in M_h .
 \end{aligned}$$

By replacing  $\delta^2 U^{n+1}$  by  $\delta^3 U^{n+1}$  everywhere in the above equation, we obtain a method which yields error estimates of the form

$$(3.7) \quad \max_{t^n} \|U^n\| \leq K_1 \{k^2 + h^{r+1}\}$$

for some positive constant  $K_1$ , using spaces with approximation properties given by (2.1). See [23] for analysis and computational discussion.

Finally we consider second-order Sobolev equations obtained by choosing  $e > 0$ ,  $c \equiv 0$ ,  $a > 0$ ,  $b > 0$ , and  $g \geq 0$  in (1.1):

$$e \frac{\partial^2 u}{\partial t^2} - \nabla \cdot (a \nabla u + b \nabla \frac{\partial u}{\partial t} + g \nabla \frac{\partial^2 u}{\partial t^2}) = f(x, t, u) .$$

Equations of this type arise in hydrodynamics and applications of viscoelasticity [6, 7, 25, 29, 30, 31, 33, 34] and numerical approximations have been

studied analytically in [20]. If  $g > 0$ , a method with four time levels is needed to obtain time-truncation errors of  $O(k)$ . This method is given by

$$\begin{aligned}
 (3.8) \quad & k^{-2} (\epsilon \delta^2 U^{n+1}, \chi) + (a \nabla U^n, \nabla \chi) + k^{-1} (b \nabla (U^{n+1} - U^{n-1}), \nabla \chi) \\
 & + k^{-2} (g \nabla \delta^2 U^{n+1}, \nabla \chi) + \frac{(kb + g)^2}{k^2 \epsilon} \left[ \left( \frac{\partial^2}{\partial x \partial y} \delta^3 U^{n+1}, \frac{\partial^2}{\partial x \partial y} \chi \right) \right. \\
 & \left. + \left( \frac{\partial^2}{\partial x \partial z} \delta^3 U^{n+1}, \frac{\partial^2}{\partial x \partial z} \chi \right) + \left( \frac{\partial^2}{\partial y \partial z} \delta^3 U^{n+1}, \frac{\partial^2}{\partial y \partial z} \chi \right) \right] \\
 & + \frac{(kb + g)^3}{k^2 \epsilon^2} \left( \frac{\partial^3}{\partial x \partial y \partial z} \delta^3 U^{n+1}, \frac{\partial^3}{\partial x \partial y \partial z} \chi \right) \\
 & = (f(t^n, U^n), \chi) \quad , \quad \chi \in M_h .
 \end{aligned}$$

Note that if  $g \geq 0$  and  $b > 0$ , the  $\delta^3 U^{n+1}$  terms in (3.8) can be replaced by  $\delta^2 U^{n+1}$  terms to obtain a three level method which yields error estimates of the form

$$(3.9) \quad \max_{t^n} \|U^n\| \leq K_1 \{k^2 + h^{r+1}\}$$

for some constant  $K_1$ . For details and analysis, see [23].

IV. DIRECT METHODS

Now that the basic AD ideas have been presented in the constant coefficient case in  $R^3$  we shall discuss methods for treating the nonlinear coefficients in (1.1) in  $R^2$ . Extensions to  $R^3$  should be obvious. We shall first consider methods which we term direct methods which have been derived from the Laplace-modified ideas presented in [12] and used extensively in [8, 11, 12, 15, 27, 28].

Again, we first consider parabolic equations with  $e \equiv b \equiv g \equiv 0$  in (1.1):

$$(4.0) \quad c(x, u) \frac{\partial u}{\partial t} - \nabla \cdot (a(x, u) \nabla u) = f(x, t, u).$$

The basic idea of direct methods is to replace the variable coefficients at the top time levels by a constant, or sequence of constants, which is "close" to the true coefficient. Then the error made by this replacement is multiplied by a "small" term obtained by extrapolations from previous time levels. Once constant coefficient values are obtained at the advanced time levels the AD procedures described in Section 3 can be applied.

Since many important problems have different-sized diffusion components in different directions, we shall not use only Laplace-modified methods but shall allow a direction-oriented modification. We then modify (3.1) as follows. Let  $c_0$ ,  $a_1$ , and  $a_2$  be fixed, let

$$(4.1) \quad \begin{aligned} a) \quad \tilde{c}^{n+1} &= c(\vec{x}, E(\mu) U^{n+1}) - c_0 \\ b) \quad \tilde{a}_1^{n+1} &= a_x(\vec{x}, E(\mu) U^{n+1}) - a_1 \\ c) \quad \tilde{a}_2^{n+1} &= a_y(\vec{x}, E(\mu) U^{n+1}) - a_2 \end{aligned}$$

where  $a_x$  and  $a_y$  are the components of the vector  $a$  and let  $U^{n+1}$  satisfy

$$\begin{aligned}
& k^{-1} (c(E(\mu) U^{n+1}) \delta U^{n+1}, x) + \beta \{ (a_x(E(\mu) U^{n+1}) \frac{\partial}{\partial x} U^{n+1}, \frac{\partial}{\partial x} x) \\
& + (a_y(E(\mu) U^{n+1}) \frac{\partial}{\partial y} U^{n+1}, \frac{\partial}{\partial y} x) \\
& + \frac{k\beta^2 a_1 a_2}{c_0} (\frac{\partial^2}{\partial x \partial y} D(\mu) U^{n+1}, \frac{\partial^2}{\partial x \partial y} x) \\
(4.2) \quad & = k^{-1} ([\tilde{c}^{n+1} F(\mu) U^{n+1} + c_0 \{ a_1 \delta U^n + a_2 \delta U^{n-1} \}], x) \\
& + \beta (\tilde{a}_1^{n+1} \frac{\partial}{\partial x} G(\mu) U^{n+1}, \frac{\partial}{\partial x} x) + \beta (\tilde{a}_2^{n+1} \frac{\partial}{\partial y} G(\mu) U^{n+1}, \frac{\partial}{\partial y} x) \\
& + \beta (f(t^{n+1}, E(\mu) U^{n+1}), x) , \quad x \in M_h .
\end{aligned}$$

The choices of  $a_i(\mu)$ ,  $i = 1, 2, 3$ ,  $\beta(\mu)$  and  $E(\mu)$  are given in Table 1 for  $\mu = 1, 2, 3$ . Choices of  $D(\mu)$ ,  $F(\mu)$  and  $G(\mu)$  are given in Table 2 for methods with time-truncation errors of order  $k^\mu$  for  $\mu = 1, 2$ , and 3. As an example, the case  $\mu = 1$ , can be written in the form

$$\begin{aligned}
& k^{-1} (c_0 \delta U^{n+1}, x) + (a_1 \frac{\partial}{\partial x} U^{n+1}, \frac{\partial}{\partial x} x) + (a_2 \frac{\partial}{\partial y} U^{n+1}, \frac{\partial}{\partial y} x) \\
& + \frac{k a_1 a_2}{c_0} (\frac{\partial^2}{\partial x \partial y} \delta U^{n+1}, \frac{\partial^2}{\partial x \partial y} x) \\
(4.3) \quad & = k^{-1} ([c(U^n) - c_0] \delta U^n, x) - ([a_x(U^n) - a_1] \frac{\partial}{\partial x} U^n, \frac{\partial}{\partial x} x) \\
& - ([a_y(U^n) - a_2] \frac{\partial}{\partial y} U^n, \frac{\partial}{\partial y} x) + (f(t^{n+1}, U^n), x) , \quad x \in M_h .
\end{aligned}$$

This equation has only constant coefficients at the advanced time level. The operator for the advanced time level can thus be factored easily into a product of two one-dimensional operators. We note that the first-order method is similar to that discussed in [11, 26, 27]. The first second-order method from Table 2 is similar to the direct method discussed in [28], which has a Crank-Nicolson base scheme, but this method is strongly stable. Both of the aforementioned methods required constraints of the form

$$(4.4) \quad \begin{aligned} a) \quad & \frac{3}{4} c(x, E(2) U^{n+1}) < c_0 < \frac{5}{4} c(x, E(2) U^{n+1}) \\ b) \quad & a(x, E(2) U^{n+1}) < a_0. \end{aligned}$$

Although this is a very mild constraint on  $a_0$  it is a fairly restrictive two-sided constraint on  $c_0$  and is noted in Table 2. Certain patch approximation techniques presented in [26, 27, 28] help to make this constraint localized and thus less restrictive. Another second-order method which has only one-sided constraints but requires an extra matrix inversion at each time step is also presented in Table 2 and has been analyzed by Bramble and the author. If  $c$  is a positive constant, we have presented two third-order direct methods. The first has two-sided constraints on  $a_1$  and  $a_2$  while the second obtains one-sided constraints at greater computational expense as before. Analysis and details will appear elsewhere. Note that the operator  $S_Y$  appearing in Table 2 is given in (3.4).

In the analysis of all the methods presented by (4.2) and Table 2, the use of backward differentiation multistep base methods and the projection  $Z = P_x P_y P_z u$  instead of the usual weighted elliptic projection allows very weak mesh-ratio conditions of the form:

$$(4.5) \quad \begin{aligned} a) \quad & h^r < k \quad , \text{ for } d = 2, \\ b) \quad & c_1 h^r < k < ch^{\frac{d}{2\mu}} \quad , \text{ for } d = 3. \end{aligned}$$

The use of this projection also requires only the same smoothness for the AD variants as for the base schemes. Use of only the elliptic projection requires more smoothness in time than the results presented here (see [3]).

Using the ideas described above, we can also define AD methods for nonlinear Sobolev equations and wave equations. For example let  $e \equiv g \equiv 0$  and  $a$ ,  $b$ , and  $c$  be uniformly bounded from below by positive constants in (1.1):

$$c(x, u) \frac{\partial u}{\partial t} - \nabla \cdot (a(x, u) \nabla u + b(x, u) \nabla \frac{\partial u}{\partial t}) = f(x, t, u).$$

We can then consider, for  $\mu = 1, 2$ ,

$$\begin{aligned}
 & k^{-1} \{c(E(\mu) U^{n+1}) \delta U^{n+1}, x\} + \beta \{(a_x(E(\mu) U^{n+1}) \frac{\partial}{\partial x} U^{n+1}, \frac{\partial}{\partial x} x) \\
 & + (a_y(E(\mu) U^{n+1}) \frac{\partial}{\partial y} U^{n+1}, \frac{\partial}{\partial y} x)\} \\
 & + k^{-1} \{(b_x(E(\mu) U^{n+1}) \frac{\partial}{\partial x} \delta U^{n+1}, \frac{\partial}{\partial x} x) \\
 & + (b_y(E(\mu) U^{n+1}) \frac{\partial}{\partial y} \delta U^{n+1}, \frac{\partial}{\partial y} x)\} \\
 & + \frac{(b_1 + k \beta a_1)(b_2 + k \beta a_2)}{kc_0} (\frac{\partial^2}{\partial x \partial y} D(\mu) U^{n+1}, \frac{\partial^2}{\partial x \partial y} x) \\
 (4.6) \quad & = k^{-1} ([\tilde{c}^{n+1} F(\mu) U^{n+1} + c_0 \alpha_1 \delta U^n], x) \\
 & + k^{-1} ([\tilde{b}_1^{n+1} \frac{\partial}{\partial x} F(\mu) U^{n+1} + b_1 \alpha_1 \frac{\partial}{\partial x} \delta U^n], \frac{\partial}{\partial x} x) \\
 & + k^{-1} ([\tilde{b}_2^{n+1} \frac{\partial}{\partial y} F(\mu) U^{n+1} + b_2 \alpha_1 \frac{\partial}{\partial y} \delta U^n], \frac{\partial}{\partial y} x) \\
 & + \beta \{(\tilde{a}_1^{n+1} \frac{\partial}{\partial x} G(\mu) U^{n+1}, \frac{\partial}{\partial x} x) + (\tilde{a}_2^{n+1} \frac{\partial}{\partial y} G(\mu) U^{n+1}, \frac{\partial}{\partial y} x)\} \\
 & + \beta (f(t^{n+1}, E(\mu) U^{n+1}), x) \quad , \quad x \in M_h ,
 \end{aligned}$$

where  $b_x, b_y, b_1, b_2, \tilde{b}_1$ , and  $\tilde{b}_2$  are analogous to the corresponding coefficients for  $a$  (see (4.1)) and  $F, D, G$ , and  $E$  are from Table 2 as before. We note that the base scheme used for time-stepping the Sobolev equation here is a backward differentiation multistep method and is different from that used for similar equations in Section 3. Corresponding direct methods could be defined from the methods of Section 3. Analysis of (4.6) will appear in [22].

In a similar manner, direct methods could be used to obtain efficient AD methods for hyperbolic and second-order Sobolev equations where  $e(x, u)$  is nonlinear in (1.1). Techniques like those used in [20] are required. Detailed descriptions and analysis of these methods will appear elsewhere.

## V. ITERATIVE METHODS

In this section we discuss iterative stabilization methods for treating the nonlinearities in the coefficients as an alternative to direct methods. We shall use the ideas developed in [14, 19] and later used for multistep methods in [5, 21]. The basic idea for the base scheme is to factor the matrix arising from the linear algebra problem at one time-step, say the initial time-step. We then use this factored matrix as a preconditioner in a preconditioned conjugate gradient iterative procedure to keep from factoring a new matrix at each time step. This factored matrix is comparable to the matrix which should be inverted at each time level. Thus we can extrapolate from past values to obtain the proper accuracy and only iterate sufficiently often to stabilize the process. For many problems this requires only two to four iterations per time step. If the coefficients begin to change considerably, one should refactor to obtain a more comparable preconditioner periodically. For discussion of these computational complexities and work estimates, see [11, 14, 19, 20, 24].

The use of iterative stabilization in conjunction with AD methods was first presented in [11]. The factored operator  $S_y$  from (3.4) was used as a preconditioner in a first-order time method. However, since the base method did not include AD perturbation terms as in (3.2), a mesh-ratio restriction of the form

$$(5.1) \quad k < K h^2 \quad , \text{ for } d = 2 ,$$

is required in [11] in order that the preconditioner be comparable to the linear operator which should be solved at each time step. Since we include an AD perturbation term in our base scheme, we obtain comparability with the preconditioner with no mesh-ratio restrictions. The only mesh-ratio restrictions required by the methods presented here are the weak conditions given by (4.5).

The base scheme for the methods to be presented in this section for parabolic problems from (4.0) is

$$\begin{aligned}
& k^{-1} (c(x, E(\mu) U^{n+1}) \delta U^{n+1}, x) \\
& + \beta (a_x(x, E(\mu) U^{n+1}) \frac{\partial}{\partial x} U^{n+1}, \frac{\partial}{\partial x} x) \\
& + \beta (a_y(x, E(\mu) U^{n+1}) \frac{\partial}{\partial y} U^{n+1}, \frac{\partial}{\partial y} x) \\
(5.2) \quad & + \frac{k \beta^2 a_1 a_2}{c_0} (\frac{\partial^2}{\partial x \partial y} D(\mu) U^{n+1}, \frac{\partial^2}{\partial x \partial y} x) \\
& = k^{-1} (c(x, E(\mu) U^{n+1}) [a_1 \delta U^n + a_2 \delta U^{n-1}], x) \\
& + \beta (f(x, t^{n+1}, E(\mu) U^{n+1}), x), \quad x \in M_h,
\end{aligned}$$

where  $a_1$ ,  $a_2$ , and  $c_0$  are as in (4.1) and  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$ ,  $E(\mu)$  and  $D(\mu)$  are as in Table 1. We shall next define our iterative stabilization schemes.

We first present the linear equations arising from (5.2) for the case  $\mu = 3$  and note that there is no direct AD factorization possible for these equations. This motivates the introduction of a fixed preconditioner for which the linear equations do have an AD factorization.

We define two orderings on the nodes in  $\Omega = [0,1]^2$ . The first is a global ordering which assigns one of the numbers  $1, 2, \dots, M$  to each node in  $\Omega$ . The second is a tensor product ordering of the  $M$  nodes. Grid lines of the form  $x = x_j$ ,  $0 < x_j < 1$ , are numbered  $1, 2, \dots, M_x$  while grid lines of the form  $y = y_j$ ,  $0 < y_j < 1$  are numbered  $1, 2, \dots, M_y$ . With each node  $i$ , we associate an  $x$ -grid line and a  $y$ -grid line. The tensor product index of the node  $i$  is the pair  $(m(i), n(i))$ , where  $m(i)$  is the index of the  $x$ -grid line and  $n(i)$  is the index of the  $y$ -grid line. We then denote the tensor product basis as

$$(5.3) \quad B_i(\vec{x}) = \phi_{m(i)}(x) \psi_{n(i)}(y) = \phi_m(x) \psi_n(y), \quad 1 \leq i \leq M,$$

where  $\{\phi_m(x)\}_{m=1}^{M_x}$  and  $\{\psi_n(y)\}_{n=1}^{M_y}$  are bases for the one-dimensional spaces  $M_h^0([0,1])$  for  $x$  or  $y$  in  $[0,1]$ , respectively.

Let  $U^P$  from (5.2) be written as

$$(5.4) \quad U^P = \sum_{i=1}^M \xi_i^P B_i(\vec{x}) = \sum_{m=1}^{M_x} \sum_{n=1}^{M_y} \xi_{mn}^P \phi_m(x) \psi_n(y).$$

Using (5.4), (5.2) with  $\mu = 3$  can be written as

$$(5.5) \quad L^{n+1} \{ \xi^{n+1} - \xi^n \} = C^n(\xi) \left\{ \sum_{j=1}^2 \alpha_j \delta \xi^{n+1-j} \right\} + k \{ F_1^n(\xi) + F_2^n(\xi) \} \\ \equiv F^n(\xi)$$

where the matrices and vectors in (5.5) are defined by

$$(5.6) \quad \begin{aligned} a) \quad L^n &= C^n + k A^n + k^2 G^n, \\ b) \quad C^n &= ((c(E(3) U^{n+1}) B_j, B_i)), \\ c) \quad A^n &= \beta ((a_x(E(3) U^{n+1}) \frac{\partial}{\partial x} B_j, \frac{\partial}{\partial x} B_i) \\ &\quad + (a_y(E(3) U^{n+1}) \frac{\partial}{\partial y} B_j, \frac{\partial}{\partial y} B_i)), \\ d) \quad G^n &= \frac{\beta^2 a_1 a_2}{c_o} ((\frac{\partial^2}{\partial x \partial y} B_j, \frac{\partial^2}{\partial x \partial y} B_i)), \\ e) \quad F_1^n(\xi) &= -A^{n+1} \xi^{n+1} + \beta ((f(t^{n+1}, E(3) U^{n+1}), B_j)), \\ f) \quad F_2^n(\xi) &= G^n [\xi^n - \xi^{n+1}], \end{aligned}$$

for  $i, j = 1, 2, \dots, M$ .

Instead of solving (5.5) exactly, we shall approximate its solution by using an iterative procedure which has been preconditioned by  $\bar{L}^0$  the matrix (5.6.a) with  $c$ ,  $a_x$ , and  $a_y$  replaced by  $c_o$ ,  $a_1$ , and  $a_2$ , respectively. Since the matrix  $\bar{L}^0$  has constant coefficients, we can use the tensor product property of the basis to factor  $\bar{L}^0$  into the product

$$(5.7) \quad \bar{L}^0 = (C_x + k A_x)(C_y + k A_y)$$

where

$$a) \quad C_x = ((c_0^{1/2} \phi_i(x), \phi_j(x)))$$

$$b) \quad A_x = (\beta a_1 c_0^{-1/2} (\phi_j(x), \phi_l(x)))$$

$$c) \quad C_y = ((c_0^{1/2} \psi_n(y), \psi_m(y)))$$

$$d) \quad A_y = (\beta a_2 c_0^{-1/2} (\psi_n(y), \psi_m(y)))$$

for  $i, j = 1, \dots, M_x$ , and  $m, n = 1, \dots, M_y$ . Thus inverting  $\bar{L}^0$  corresponds to solving two one-dimensional problems successively.

The preconditioning process eliminates the need for factoring new matrices at each time step and reduces the problem to successive solution of one-dimensional problems, while the iterative procedure stabilizes the resulting problem. The stabilization process requires iteration only until a predetermined norm reduction is achieved.

Denote by

$$(5.8) \quad V^S = \sum_{i=1}^{M_x} \theta_i^S B_i(x) = \sum_{m=1}^{M_y} \sum_{n=1}^{M_y} \theta_{mn}^S \phi_m(x) \psi_n(y),$$

the approximation to  $U^S$  produced by only approximately solving (5.5) using  $\bar{L}^0$ . Assume sufficiently accurate starting values have been obtained (see [3,4]). Assuming  $V^0, \dots, V^n$  have been determined, we shall determine the  $M$ -dimensional vector  $\theta^{n+1}$  (and thus  $V^{n+1}$  from (5.8)) using a preconditioned iterative method to approximate  $\xi^{n+1}$  from (5.5). As an initial guess for  $\xi^{n+1} - \xi^n$ , we shall extrapolate from previously determined values. Specifically, for the method under consideration having time-truncation error  $O(k^3)$ , we shall use as an initialization for our iterative procedure

$$(5.9) \quad x_o = (\theta^{n+1} - \theta^n) - \delta^4 \theta^{n+1}.$$

Since we are using previously determined  $\theta^i$  in the matrix problem (5.5) to determine  $\theta^{n+1}$ , our errors accumulate.

In order to analyze the cumulative error, we first consider the single step error. We define  $\bar{\theta}^{n+1}$  to satisfy

$$(5.10) \quad L^{n+1} \{ \bar{\theta}^{n+1} - \theta^n \} = F^n(\theta), \quad \text{for } n \geq 3.$$

Thus  $\bar{\theta}^{n+1}$  would be the exact solution of (5.5) if the computed values of  $\theta^k$  from previous approximate solutions of (5.5) using  $L_o$  had been used for  $k \leq n$ . We can use any preconditioned iterative method which yields norm reductions of the form

$$(5.11) \quad \| (L^{n+1})^{1/2} (\bar{\theta}^{n+1} - \theta^{n+1}) \|_e$$

$$\leq \rho_n \| (L^{n+1})^{1/2} (\bar{\theta}^{n+1} - \theta^{n+1} + \delta^4 \theta^{n+1}) \|_e$$

where  $0 < \rho_n < 1$  and the subscript e denotes the Euclidean norm of the vector. A specific iterative procedure for obtaining (4.8) is the preconditioned conjugate gradient method analyzed in [1, 2, 13, 14, 19].

Then, letting

$$(5.12) \quad \bar{v}^s = \sum_{i=1}^M \bar{\theta}_i^s B_i(\vec{x}) = \sum_{m=1}^{M_x} \sum_{n=1}^{M_y} \bar{\theta}_{mn}^s \phi_m(x) \psi_n(y),$$

with  $\bar{\theta}^s$  defined in (5.10), we see that  $v^{n+1}$  and  $\bar{v}^{n+1}$  satisfy

$$\begin{aligned}
& k^{-1} (c(x, E(\mu) v^{n+1}) \delta v^{n+1}, x) \\
& + \beta (a_x(x, E(\mu) v^{n+1}) \frac{\partial}{\partial x} v^{n+1}, \frac{\partial}{\partial x} x) \\
& + \beta (a_y(x, E(\mu) v^{n+1}) \frac{\partial}{\partial y} v^{n+1}, \frac{\partial}{\partial y} x) \\
& + \frac{k \beta^2 a_1 a_2}{c_0} (\frac{\partial^2}{\partial x \partial y} D(\mu) v^{n+1}, \frac{\partial^2}{\partial x \partial y} x) \\
& = k^{-1} (c(x, E(\mu) v^{n+1}) [\alpha_1 \delta v^n + \alpha_2 \delta v^{n-1}], x) \\
(5.13) \quad & + \beta (f(x, t^{n+1}, E(\mu) v^{n+1}), x) \\
& + k^{-1} (c(x, E(\mu) v^{n+1}) [v^{n+1} - \bar{v}^{n+1}], x) \\
& + \beta (a_x(x, E(\mu) v^{n+1}) \frac{\partial}{\partial x} (v^{n+1} - \bar{v}^{n+1}), \frac{\partial}{\partial x} x) \\
& + \beta (a_y(x, E(\mu) v^{n+1}) \frac{\partial}{\partial y} (v^{n+1} - \bar{v}^{n+1}), \frac{\partial}{\partial y} x) \\
& + \frac{k \beta^2 a_1 a_2}{c_0} (\frac{\partial^2}{\partial x \partial y} [v^{n+1} - \bar{v}^{n+1}], \frac{\partial^2}{\partial x \partial y} x), \quad x \in M_h,
\end{aligned}$$

where the last four terms measure the single step error arising from the iterative stabilization. We must iterate only sufficiently often to control these terms in the analysis. Since  $\bar{L}_0$  is a sequence of one-dimensional operators, we can very efficiently update  $\bar{L}_0$  if  $L_n$  drifts far away from  $\bar{L}_0$ . Analysis and details will appear in [3].

Note that in preconditioned iterative methods, only the preconditioner is inverted. In this case, that is only a sequence of one-dimensional problems. If the basis functions in the one-dimensional problem are linear (tensor products of linears for the basis for  $M_h$ ) the matrices to be inverted are tridiagonal and if the basis functions are quadratic the matrices are pentadiagonal. Thus if  $d = 2$  or  $3$  the work estimate is  $O(M_x M_y)$  or  $O(M_x M_y M_z)$ ,

respectively. Thus the work is proportional to the total number of unknowns in the problem and optimal order work estimates are obtained (see [11, 14, 24, 27, 28]).

The storage requirements are also very attractive for AD methods. Since the matrix problem is treated as a series of one-dimensional problems, only the data corresponding to one grid line are required in core at any given time. In two dimensions the storage requirements for these AD methods are comparable to those of a frontal elimination solver, but these methods require considerably less I/O. In three dimensions the frontal elimination solvers require that a plane of data be in core, while these methods only require one line of data. Clearly all of the above remarks apply to each of the AD methods presented here, not only to the iterative variants.

The author has applied iterative stabilization methods to problems of hyperbolic and Sobolev types in [19, 20]. The extension of these iterative ideas to AD methods for equations of these types follows from the ideas presented above for parabolic problems.

TABLE 1: BACKWARD DIFFERENTIATION MULTISTEP METHODS

$\mu$	$\beta(\mu)$	$\alpha_1(\mu)$	$\alpha_2(\mu)$	$\alpha_3(\mu)$	$E(\mu) U^{n+1}$
1	1	0	0	0	$U^{n+1} - \delta U^{n+1}$
2	2/3	1/3	0	0	$U^{n+1} - \delta^2 U^{n+1}$
3	6/11	7/11	-2/11	0	$U^{n+1} - \delta^3 U^{n+1}$
4	12/25	23/25	-13/25	3/25	$U^{n+1} - \delta^4 U^{n+1}$

TABLE 2: DIRECT METHODS

$\mu$	$D(\mu) U^{n+1}$	$F(\mu) U^{n+1}$	$G(\mu) U^{n+1}$	Coefficient Constraints
1	$\delta U^{n+1}$	$\delta^2 U^{n+1}$	$\delta U^{n+1}$	one-sided ( $c_o$ )
2	$\delta U^{n+1}$	$\delta^3 U^{n+1}$	$\delta^2 U^{n+1}$	two-sided ( $c_o$ )
2	$\delta U^{n+1}$	$\delta^2 U^{n+1} - k c_o S_Y^{-1} \delta [c_n^{-1} L_n U^n]$	$\delta^2 U^{n+1}$	one-sided ( $c_o$ )
3	$\delta^2 U^{n+1}$	—	$\delta^3 U^{n+1}$	two-sided ( $a_o$ )
3	$\delta U^{n+1} - c_o S_Y^{-1} U^n$	—	$\delta^2 U^{n+1} - c_o S_Y^{-1} \delta^2 U^n$	one-sided ( $a_o$ )

REFERENCES

1. O. Axelsson, "On Preconditioning and Convergence Acceleration in Sparse Matrix Problems", CERN European Organization for Nuclear Research, Geneva, 1974.
2. O. Axelsson, "On the Computational Complexity of Some Matrix Iterative Algorithms", Report 74.06, Dept. of Computer Science, Chalmers University of Technology, Goteborg, 1974.
3. J. H. Bramble and R. E. Ewing, "Alternating Direction Multistep Methods for Parabolic Problems - Iterative Stabilization" (in preparation).
4. J. H. Bramble and R. E. Ewing, "Efficient Starting Procedures for High-Order Time-Stepping Methods for Differential Equations", (in preparation).
5. J. H. Bramble and P. H. Sammon, "Efficient Higher-Order Multistep Methods for Parabolic Problems" (in preparation).
6. B. D. Coleman and W. Noll, "The Thermodynamics of Elastic Materials with Heat conduction and Viscosity", Arch. Rat. Mech. Anal. 13 (1963), pp. 167-178.
7. C. M. Dafermos, "The Mixed Initial-Boundary Problem for Equations of Nonlinear Viscoelasticity", J. of Diff. Eqn. 6 (1969), pp. 71-81.
8. J. E. Dendy, Jr., "An Analysis of some Galerkin Schemes for the Solution of Nonlinear Time-Dependent Problems", SIAM J. Numer. Anal. 12 (1975), pp. 541-565.
9. J. E. Dendy, Jr., and G. Fairweather, "Alternating-Direction Galerkin Methods for Parabolic and Hyperbolic Problems on Rectangular Polygons", SIAM J. Numer. Anal. 12 (1975), pp. 144-163.

10. J. Douglas, Jr., "Alternating-Direction Methods for Three Space Variables", *Numerische Mathematik* 4 (1962), pp. 41-63.
11. J. Douglas, Jr., "Effective Time-Stepping Methods for the Numerical Solution of Nonlinear Parabolic Problems", to appear in the Proceedings of MAFELAP78, Brunel University, 1978.
12. J. Douglas and T. Dupont, "Alternating-Direction Galerkin Methods on Rectangles", in Proceedings Symposium on Numerical Solution of Partial Differential Equations, II, B. Hubbard (ed.), Academic Press, New York, 1971, pp. 133-214.
13. J. Douglas, Jr., and T. Dupont, "Preconditioned Conjugate Gradient Iteration Applied to Galerkin Methods for a Mildly Nonlinear Dirichlet Problem", Sparse Matrix Calculations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 333-348.
14. J. Douglas, Jr., T. Dupont, and R. E. Ewing, "Incomplete Iteration for Time-Stepping a Galerkin Method for a Quasilinear Parabolic Problem", *SIAM J. Numer. Anal.* 16 (1979), pp. 503-522.
15. J. Douglas, T. Dupont, and P. Percell, "A Time-Stepping Method for Galerkin Approximations for Nonlinear Parabolic Equations", to appear in Proceedings of the Conference on Numerical Analysis held at Dundee, Scotland, June 28-July 1, 1977.
16. J. Douglas, Jr., T. Dupont, and L. Wahlbin, "Optimal  $L_\infty$  Error Estimates for Galerkin Approximations to Solutions of Two-Point Boundary Value Problems", *Math. Comp.* 29 (1975), pp. 475-483.
17. J. Douglas, Jr., and J. E. Gunn, "A General Formulation of Alternating-Direction Methods", *Numer. Math.* 6 (1964), pp. 428-453.
18. J. Douglas, Jr., and H. H. Rachford, "On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables", *Trans. Am. Math. Soc.* 82 (1956), pp. 421-439.

19. R. E. Ewing, "Time-Stepping Galerkin Methods for Nonlinear Sobolev Partial Differential Equations", SIAM J. Numer. Anal. 15 (1978), pp. 1125-1150.
20. R. E. Ewing, "On Efficient Time-Stepping Methods for Nonlinear Partial Differential Equations, Math. Res. Cent. Rep. #1996, University of Wisconsin, Madison (1979), and Computers and Math. with Appl. 6 (1980), pp. 1-13.
21. R. E. Ewing, "Efficient Multistep Procedures for Nonlinear Parabolic Problems with Nonlinear Neumann Boundary Conditions, Math. Res. Cent. Rep. #1982, University of Wisconsin, Madison (1979), and Trans. 25th Conf. Army Math. ARO Report 80-1, and Calcolo (to appear).
22. R. E. Ewing and L. J. Hayes, "Alternating Direction Galerkin Methods for Nonlinear Sobolev Partial Differential Equations", (in preparation).
23. R. E. Ewing and L. J. Hayes, "Alternating Direction Galerkin Methods for Some Third and Fourth Order Partial Differential Equations", (in preparation).
24. R. E. Ewing and T. R. Russell, "Efficient Time-Stepping Methods for Miscible Displacement Problems in Porous Media", SIAM J. Numer. Anal. (to appear).
25. C. M. Greenberg, R. C. MacCamy, and V. J. Mizel, "On the Existence, Uniqueness, and Stability of Solutions of the Equation  $\sigma^2(u_x)^2 u_{xx} + \lambda u_{xxx} = \rho_0 u_{ttt}$ ", J. Math. and Mech. 17 (1968), pp. 707-728.
26. L. J. Hayes, "Generalization of Galerkin Alternating-Direction Methods to Nonrectangular Regions Using Isoparametric Elements", Ph.D. Thesis, University of Texas at Austin, December 1977.

27. L. J. Hayes, "Generalization of Galerkin Alternating-Direction Methods to Nonrectangular Regions Using Patch Approximations", SIAM J. Numer. Anal. (to appear).
28. L. J. Hayes and P. Percell, "Efficient Second-Order Time-Stepping Procedures for Nonlinear Parabolic Equations", SIAM J. Numer. Anal. (to appear).
29. M. Lighthill, "Dynamics of Rotating Fluids: a Survey", J. Fluid Mech. 26 (1966), pp. 411-436.
30. A. E. H. Love, A Treatise on the Mathematical Theory of Elasticity, Dover, New York, 1944.
31. R. C. MacCamy, "Existence, Uniqueness, and Stability of Solutions of the Equation  $u_{tt} = \frac{\partial}{\partial x} (\sigma(u_x) + \lambda(u_x)u_{xt})$ ", Report 68-18, Dept. of Math., Carnegie Inst. of Technology, Carnegie-Mellon University.
32. D. W. Peaceman and H. H. Rachford, "The Numerical Solution of Parabolic and Elliptic Differential Equations", SIAM J. 3 (1955), pp. 28-41.
33. G. W. Platzman, "The Eigenvalues of Laplace's Tidal Equations", Quart. J. of Royal Meteorological Soc. 94 (1968), pp. 225-248.
34. R. E. Showalter, "Regularization and Approximation of Second-Order Evolution Equations", SIAM J. Math. Anal. 7 (1976), pp. 461-472.

GALERKIN METHODS FOR MISCIBLE DISPLACEMENT  
PROBLEMS WITH POINT SOURCES AND SINKS -  
UNIT MOBILITY RATIO CASE

by

Richard E. Ewing  
Mobil Research and Development Corporation  
P. O. Box 900  
Dallas, Texas 75221

and

Mary Fanett Wheeler  
Department of Mathematics  
Rice University  
Houston, Texas 77001

Lecture by Mary Fanett Wheeler.

GALERKIN METHODS FOR MISCIBLE DISPLACEMENT  
PROBLEMS WITH POINT SOURCES AND SINKS -  
UNIT MOBILITY RATIO CASE

1. Introduction

In [7] the authors presented and analyzed certain numerical approximations by Galerkin methods for a problem arising in the miscible displacement of one incompressible fluid by another in a porous medium. Extensions of these methods to more efficient time-stepping procedures and more general boundary conditions [8], to interior penalty procedures [17], to methods of characteristics [13], to self-adaptive simulation techniques [6], and to mixed methods for pressure [5] have since been developed. These analyses were surveyed in [3]. All of the above analyses have made a major, and probably unphysical, assumption that the sources and sinks were smoothly distributed and the resulting functions of interest were thus fairly smooth in space. In this paper we shall present the first convergence analysis on this problem to appear in the literature where the sources and sinks are considered as point singularities or Dirac measures. The resulting pressure function thus has a finite number of logarithmic singularities located at the various wells. Since the resulting functions are considerably less smooth than previously assumed, the convergence rates obtained in this paper are slower than those previously obtained.

At present we are only able to analyze the special case where the viscosity of the invading fluid is equal to the viscosity of the resident fluid. In this case, the mobility ratio (see [5, 7]) is equal to one and the equation for pressure is a linear equation and can be uncoupled from the concentration equation and solved once for all time. Analysis for the general case when there is a nonlinear coupling between the pressure and concentration equations is in process and will appear elsewhere.

A set of model equations for our physical problem is given next. For a more detailed description of the physical problem, see [7, 11, 15]. Find the concentration  $c = c(x, t)$  and  $p = p(x, t)$  satisfying the following set of equations:

$$(1.1) \quad a) \quad \nabla \cdot [a(x)\{\nabla p - \gamma \nabla g\}] = -\nabla \cdot u = \sum_{j=1}^N Q_j(t) \delta(x - x_j), \quad x \in \Omega, t \in J,$$

$$b) \quad \phi \frac{\partial c}{\partial t} + u \cdot \nabla c = \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial}{\partial x_i} [D_{ij}(x, u) \frac{\partial}{\partial x_j} c]$$

$$= \sum_{j=1}^N Q_j(t) (\tilde{c} - c) \delta(x - x_j), \quad x \in \Omega, t \in J,$$

with an initial condition and no flow boundary conditions given by

$$(1.2) \quad a) \quad c(x, 0) = c_0(x), \quad x \in \Omega,$$

$$b) \quad u \cdot v = 0, \quad x \in \partial\Omega, t \in J,$$

$$c) \quad \frac{\partial c}{\partial v} = 0, \quad x \in \partial\Omega, t \in J,$$

where  $\Omega$  is a bounded domain in  $\mathbb{R}^2$ ,  $J = [0, T]$ , and  $v$  is the outward unit normal vector on  $\partial\Omega$ , the boundary of  $\Omega$ . Here  $a = a(x)$ ,  $\gamma = \gamma(x)$ ,  $g = g(x)$ ,  $\phi = \phi(x)$  are specified reservoir and fluid properties,  $u$  is the Darcy velocity of the fluid,  $\tilde{c}$  is the specified concentration at injection wells and the resident concentration at production wells,  $\delta(x - x_j)$  is a Dirac delta function at  $x = x_j$ , and  $Q_j(t)$  are the specified flow rates of the wells with the convention

$$(1.3) \quad a) \quad Q_j(t) > 0 \text{ for } j = 1, \dots, N/2 \quad (\text{injection wells}),$$

$$b) \quad Q_j(t) < 0 \text{ for } j = N/2 + 1, \dots, N \quad (\text{production wells}).$$

The diffusion tensor  $D_{ij}$  is given by

$$D(x, u) = (D_{ij}(x, u))$$

$$(1.4) \quad = \phi(x) D_0(x) I + \frac{\alpha_\ell}{|u|} \begin{pmatrix} u_1^2 & u_1 u_2 \\ u_1 u_2 & u_2^2 \end{pmatrix} + \frac{\alpha_t}{|u|} \begin{pmatrix} u_2^2 & -u_1 u_2 \\ -u_1 u_2 & u_1^2 \end{pmatrix}$$

where  $\alpha_\ell$  and  $\alpha_t$ , the magnitudes of longitudinal and transverse dispersion, are given constants. Here for  $v \in \mathbb{R}^2$ ,  $|v|$  is the standard Euclidean norm of the vector. We make the physically realistic assumption on  $D_0$ ,  $\alpha_\ell$ , and  $\alpha_t$  that

$$(1.5) \quad 0 < D_* |\xi|^2 < \xi^T D(x, q) \xi, \quad q \in \mathbb{R}^2, \quad \xi \in \mathbb{R}^2.$$

This gives us a coercivity property for the parabolic equation and an assumption of non-trivial diffusion and dispersion in the problem. We shall consider two separate cases for the diffusion tensor. In Case I  $\alpha_\ell = \alpha_t = 0$  and we have only molecular diffusion while in Case II  $\alpha_\ell > 0$  and  $\alpha_t > 0$  help to model physical dispersion or mixing due to the flowing motion. Results for Cases I and II are presented in Theorems 3.1 and 3.2, respectively.

Using (1.1.a), we can see that pressure can be separated into its logarithmic singularity components and a smoother component,  $\tilde{p}$ , as follows:

$$(1.6) \quad p(x, t) = \sum_{i=1}^N \frac{Q_i(t)}{2\pi} \frac{1}{a(x_i)} \ln |x - x_i| + \tilde{p}(x, t).$$

Similarly we can decompose the Darcy velocity as follows:

$$(1.7) \quad u(x, t) = \sum_{i=1}^N \frac{Q_i(t)}{2\pi} \frac{a(x)}{a(x_i)} \nabla \ln |x - x_i| + a(x) \nabla \tilde{p}.$$

We shall be more explicit about the smoothness of  $\tilde{p}$  in Section 2 after we have presented the necessary terminology (see (2.8)). The fact that pressure is assumed to have logarithmic singularities also effects the smoothness of

the concentration of the invading fluid. In particular, according to [14],

$$\int_0^T \int_{\Omega} \frac{\partial c}{\partial t}^2 dx dt \text{ is not even bounded under the point sources assumption.}$$

Therefore the convergence analysis presented in this paper is non-standard and much more difficult than for the case of smoothly distributed sources and sinks.

The paper contains two additional sections. In Section 2, terminology is developed, basic regularity and boundedness assumptions are presented, basic projections needed for the analysis are considered, and the continuous-time Galerkin approximations of (1.1) and (1.2) are defined. In Section 3, a priori error estimates for the continuous-time approximations are obtained.

$L^2$  rates of convergence for Cases I and II are given by  $h^{1-\epsilon}$  and  $h^{1/2-\epsilon}$ , respectively.

## 2. Preliminaries and Description of the Galerkin Approximations

Let  $(u, v) = \int_{\Omega} uv dx$  and  $\|u\|^2 = (u, u)$  be the standard  $L^2$  inner-product and norm. Let  $W_s^k(\Omega)$  be the Sobolev space on  $\Omega$  with norm

$$(2.1) \quad \|\psi\|_{W_s^k} = \left[ \sum_{|\alpha| \leq k} \left\| \frac{\partial^\alpha \psi}{\partial x^\alpha} \right\|_{L^s(\Omega)}^s \right]^{1/s},$$

with the usual modification for  $s = \infty$ . If  $\nabla F = (F_1, F_2)$ , write  $\|\nabla F\|_{W_s^k}$  in place of  $(\|F_1\|_{W_s^k}^s + \|F_2\|_{W_s^k}^s)^{1/s}$ . When  $s = 2$ , denote  $\|\psi\|_{W_2^k} \equiv \|\psi\|_{H^k} \equiv \|\psi\|_k$ . If  $k = 0$ ,  $\|\psi\|_0 \equiv \|\psi\|$ .

Let  $\{M_h\}$  be a family of finite-dimensional subspaces of  $H^1(\Omega)$  with the following property:

For  $p = 2$  or  $\infty$ , there exist an integer  $r > 2$  and a constant  $K_0$  such that, for  $1 \leq q \leq r$  and  $\psi \in W_p^q(\Omega)$ ,

a)  $\inf_{x \in M_h} \{ \| \psi - x \|_{W_p^0} + h \| \psi - x \|_{W_p^1} \} \leq K_0 \| \psi \|_{W_p^q} h^q,$

(2.2)

b)  $\inf_{x \in M_h} \{ \| \psi - x \|_{W_\infty^0} + h \| \psi - x \|_{W_p^1} \} \leq K_0 \| \psi \|_q h^{q-1}.$

We also define a family of finite-dimensional subspaces of  $H^1(\Omega)$  called  $\{N_h\}$  which satisfies the same property as  $\{M_h\}$  with  $r$  replaced by  $s$ . We also assume that the families  $\{M_h\}$  and  $\{N_h\}$  satisfy the following so-called "inverse hypotheses":

if  $\psi \in M_h$  or  $N_h$ , for some  $K_1 > 0$ ,

a)  $\| \psi \|_{L^p} \leq K_1 h^{\frac{2}{p}-1} \| \psi \|, \quad 2 \leq p \leq \infty$

(2.3)

b)  $\| \psi \|_1 \leq K_1 h^{-1} \| \psi \|.$

We shall use  $M_h$  to approximate  $c$  and  $N_h$  to approximate the non-logarithmic part of  $p$ .

We shall make the same boundedness assumptions and somewhat weaker smoothness assumptions on the coefficients than were made in [7, 8, 17]. We consider spaces of the form

$$\| \psi \|_{W_p^q((a,b), X)} = \{ \psi : (a,b) \rightarrow X \mid \| \frac{\partial^\alpha \psi}{\partial t^\alpha}(t) \|_X \in L^p((a,b)) \}$$

with norm

$$(2.4) \quad \| \psi \|_{W_p^q((a,b), X)} = \left[ \sum_{|\alpha| \leq q} \| \frac{\partial^\alpha \psi}{\partial t^\alpha}(t) \|_X^p \right]^{1/p},$$

where  $1 < p, q < \infty$  and  $X$  is a Sobolev space in our applications. When  $(a,b) = J$ , we shall suppress  $(a,b)$  in our notation in (2.4). Let  $(p,c)$ , the solution of (1.1)-(1.2), satisfy the following regularity assumptions:

(2.5)

- $\|c\|_{L^2(H^{2-\epsilon})} + \|c\|_{L^2(W_p^1)} + \|c\|_{L^\infty(H^{1-\epsilon})} \leq K_2 ,$
- $\|p\|_{L^\infty(H^{1-\epsilon})} \leq K_2 ,$
- $\|u\|_{L^\infty(L^{2-\epsilon})} \leq K_2 ,$
- $\|\frac{\partial c}{\partial t}\|_{L^2(L^{2-\epsilon})} \leq K_2 ,$

where  $\epsilon > 0$  can be chosen arbitrarily small,  $p$  can be chosen arbitrarily large,  $K_2 > 0$  is a fixed constant, and  $J$  has been suppressed in the index notation of the norms. These regularity assumptions are based on analysis by Sammon [14].

In our analysis we shall use two different approximations for  $c$  from  $M_h$ . We first define the  $L^2$  projection  $\hat{c}$  of  $c$  into  $M_h$  by

(2.6)

- $(\phi(c - \hat{c}), x) = 0 , \quad x \in M_h , \text{ or}$
- $(\phi(\frac{\partial c}{\partial t} - \frac{\partial \hat{c}}{\partial t}), x) = 0 , \quad x \in M_h .$

We are led to use the  $L^2$  projection of  $c$  into  $M_h$  instead of the now more standard  $H^1$  projection due to smoothness restrictions on  $c$ . Since we assume that  $\frac{\partial c}{\partial t}$  is only in  $L^2(J, L^{2-\epsilon})$  for  $\epsilon$  arbitrarily small, we are not able to treat terms like  $\frac{\partial}{\partial t}(c - \hat{c})$  in a normal fashion. Thus we have used  $\hat{c}$  to project this problem away as in (2.6.b). This causes reduced accuracy in terms like  $\nabla(c - \hat{c})$ , but the loss of accuracy was inevitable in any case due to the logarithmic singularities in pressure. We also denote by  $c_i$  the interpolant of  $c$  in  $M_h$ . We then use (2.3) and the theory of interpolation spaces to obtain the following approximation theory results:

Lemma 2.1: There exists a positive constant  $K_2 = K_2(\Omega, K_0, \epsilon)$  such that, for each  $t \in J$  and  $\epsilon$  arbitrarily small,

a)  $\|c - \hat{c}\| + h\|c - \hat{c}\|_1 \leq K_2 \|c\|_{q_1} h^{q_1}, \quad 0 < q_1 < 2 - \epsilon,$

(2.7) b)  $\|c - \hat{c}\|_{L^\infty} \leq K_2 \|c\|_{W_\infty^{q_2}} h^{q_2}, \quad 0 < q_2 < 1 - \epsilon,$

c)  $\|c - c_1\| + h[\|c - c_1\|_1 + \|c - c_1\|_{L^\infty}] \leq K_2 \|c\|_{q_3} h^{q_3}, \quad 1 < q_3 < 2 - \epsilon.$

Proof: See [2, 4, 16].

We next note that from [9] we know that  $\tilde{p}$  defined in (1.6) satisfies

$$(2.8) \quad \|\tilde{p}\|_{H^{2-\epsilon}} \leq K_3$$

for  $K_3 > 0$  and some arbitrarily small  $\epsilon > 0$ . We shall use the logarithmic part of  $p$  defined in (1.6) to form the leading part of our approximation  $P$  of  $p$ . We define  $P$  to be

$$(2.9) \quad P(x, t) = \sum_{i=1}^N \frac{Q_i(t)}{a(x_i)} \ln |x - x_i| + \tilde{P},$$

where  $x_i$ ,  $i=1, \dots, N$ , are the locations of the injection and production wells and  $\tilde{P} \in N_h$  is an approximation to  $\tilde{p}$  from (1.6) defined for each  $t \in J$  by

$$(2.10) \quad (a(\cdot) \nabla(\tilde{p} - \tilde{P}), \nabla x) = 0, \quad x \in M_h.$$

This is an example of a weighted elliptic projection used by Wheeler in [16]. We obtain the following result.

Lemma 2.2: There exists a positive constant  $K_4 = K_4(\Omega, K_0, K_2)$  such that, for each  $t \in J$  and  $\epsilon$  arbitrarily small,

$$(2.11) \quad \|\nabla(\tilde{p} - \tilde{P})\| \leq K_4 \|p\|_{H^q} h^{q-1},$$

for  $1 < q < 2 - \epsilon$ .

If we then define  $u$  and  $U$  by,

$$(2.12) \quad \begin{aligned} a) \quad u &= a(x) \cdot \nabla p = \sum_{j=1}^N Q_j(t) \frac{a(x)}{a(x_j)} \nabla \ln|x - x_j| + a(x) \cdot \nabla \tilde{p}, \\ b) \quad U &= a(x) \cdot \nabla P = \sum_{j=1}^N Q_j(t) \frac{a(x)}{a(x_j)} \nabla \ln|x - x_j| + a(x) \cdot \nabla \tilde{P}, \end{aligned}$$

we can immediately use (2.8) and (2.11) to obtain for each  $t \in J$  and  $K_5 = K_5(K_3, K_4, a(x))$ ,

$$(2.13) \quad \|u - U\| \leq K_5 h^{1-\epsilon}$$

where  $\epsilon > 0$  can be made arbitrarily small.

We next define the continuous-time approximation of  $c$  as follows: let  $C: [0, T] \rightarrow M_h$  be defined by

$$(2.14) \quad \begin{aligned} a) \quad (\phi \frac{\partial C}{\partial t}, x) + \sum_{i=1}^2 \sum_{j=1}^2 (D_{ij}(U) \frac{\partial}{\partial x_j} C, \frac{\partial}{\partial x_i} x) + (U \cdot \nabla C, x) \\ = \sum_{j=1}^{N/2} Q_j(t) (\tilde{c} - c)(x_j, t) x(x_j), \quad x \in M_h, \\ b) \quad (C(0) - c_0, x) = 0, \quad x \in M_h, \end{aligned}$$

where  $U$ ,  $P$ , and  $\tilde{P}$  are defined by (2.12.b), (2.9), and (2.10), respectively. The main results of this paper are a priori estimates for the error in approximating  $c$  from (1.1) by  $C$  from (2.14). These will appear in the next section.

### 3. A Priori Error Estimates

In this section, we shall obtain a priori bounds for the error in the concentration approximation  $C - c$ , to go with the bound of the error in the Darcy velocity approximation given by (2.13). We shall split our a priori estimates into two cases. Case I will reflect the assumption that the only diffusion present in the model is molecular diffusion and  $\alpha_\ell = \alpha_\tau = 0$  in (1.4). Case II will extend the estimates to the more difficult case of tensorial physical dispersion given by (1.4) with  $\alpha_\ell > 0$  and  $\alpha_\tau > 0$ . As expected, we obtain a reduced convergence rate for the more difficult case.

Theorem 3.1 Let  $(c,p)$  satisfy (1.1)-(1.2) and  $(C,P)$  satisfy (2.9), (2.10), and (2.14). For the molecular diffusion case, let  $\alpha_\ell = \alpha_\tau = 0$  in (1.4). There exist positive constants  $K_6 = K_6(\Omega, K_i, i=0 \dots, 5)$  and  $h_0$  such that, if  $h < h_0$ ,

$$(3.1) \quad \begin{aligned} \|c - C\|_{L^\infty(J, L^2)} + \|\nabla(c - C)\|_{L^2(J, L^2)} \\ + \left\{ \sum_{j=1}^N \int_0^T |Q_j(t)| (C - c)^2(x_j, t) dt \right\}^{1/2} \leq K_6 h^{1-\hat{\epsilon}}. \end{aligned}$$

( $\hat{\epsilon} = \hat{\epsilon}(\epsilon) > 0$  is defined in (3.19) below and can be taken arbitrarily small).

Proof: Let  $\xi = C - \hat{c}$  and  $\eta = c - \hat{c}$  with  $\hat{c}$  from (2.6) and  $C$  from (2.14). Subtract (2.6) from (2.14) and let  $x = \xi$  to obtain

$$\begin{aligned}
 & (\phi \frac{\partial \xi}{\partial t}, \xi) + (\phi D_0 \nabla \xi, \nabla \xi) + (u \cdot \nabla \xi, \xi) \\
 (3.2) \quad & = (\phi D_0 \nabla n, \nabla \xi) + (u \cdot \nabla n, \xi) + ((u - U) \cdot \nabla C, \xi) \\
 & + \sum_{j=1}^{N/2} Q_j(t)(c - C)(x_j, t) \xi(x_j, t) .
 \end{aligned}$$

For the last term on the left-hand side of (3.2), we integrate by parts (note that  $\frac{\partial u}{\partial v} = 0$  on  $\partial\Omega$ ) and use (1.1.a) with  $x = \xi$  to obtain

$$\begin{aligned}
 (u \cdot \nabla \xi, \xi) & = (u \cdot \nabla \frac{\xi^2}{2}, 1) \\
 & = - (\nabla \cdot u, \frac{\xi^2}{2}) \\
 (3.3) \quad & = - \frac{1}{2} \sum_{j=1}^N Q_j(t) \xi^2(x_j, t) \\
 & = - \frac{1}{2} \sum_{j=1}^{N/2} |Q_j(t)| \xi^2(x_j, t) + \frac{1}{2} \sum_{j=N/2+1}^N |Q_j(t)| \xi^2(x_j, t).
 \end{aligned}$$

We then combine part of the last term on the right side of (3.2) with (3.3) and replace  $c$  by  $c_j$  at the wells, to obtain

$$\begin{aligned}
 & \frac{1}{2} \frac{d}{dt} \|\phi^{1/2} \xi\|^2 + \|(\phi D_0)^{1/2} \nabla \xi\|^2 + \frac{1}{2} \sum_{j=1}^N |Q_j(t)| \xi^2(x_j, t) \\
 & = (\phi D_0 \nabla n, \nabla \xi) + (u \cdot \nabla n, \xi) + ((u - U) \cdot \nabla C, \xi) \\
 (3.4) \quad & + \sum_{j=1}^{N/2} Q_j(t) (c_j - \bar{c})(x_j, t) \xi(x_j, t) \\
 & = T_1 + T_2 + T_3 + T_4 .
 \end{aligned}$$

We next integrate (3.4) termwise on  $\tau$  in  $J_t = [0, t]$  for  $t \in J$ . The left-hand side of the resulting equation is then bounded below as follows

$$\begin{aligned}
 & \frac{1}{2} \int_0^t \frac{d}{d\tau} \|\phi^{1/2} \xi\|^2 d\tau + \int_0^t \|(\phi D_0)^{1/2} \nabla \xi\|^2 d\tau \\
 (3.5) \quad & + \frac{1}{2} \sum_{j=1}^N \int_0^t |Q_j(\tau)| \xi^2(x_j, \tau) d\tau \\
 & \geq \alpha [\|\xi(t)\|^2 + \|\nabla \xi\|_{L^2(J_t, L^2)}^2 + \sum_{j=1}^N \int_0^t |Q_j(\tau)| \xi^2(x_j, \tau) d\tau]
 \end{aligned}$$

where  $\alpha$  depends upon uniform lower bounds for the coefficients  $\phi$  and  $D_0$  such as  $D_*$  from (1.5). We next consider bounds for the terms on the right-hand side of the integrated analogue of (3.4). We note that from (2.7.a), we obtain

$$\begin{aligned}
 & \left| \int_0^t T_1 d\tau \right| \leq K \int_0^t \|\nabla n\| \|\nabla \xi\| d\tau \leq K h^{1-\epsilon} \|\nabla \xi\|_{L^2(J_t, L^2)} \\
 (3.6) \quad & \leq \frac{\alpha}{8} \|\nabla \xi\|_{L^2(J_t, L^2)}^2 + K h^{2(1-\epsilon)} ,
 \end{aligned}$$

where  $\alpha$  is from (3.5) and  $K$  is used here and in the following as a generic positive constant, usually of different size with each use. Then using

(2.3.a), (2.5.c), (2.7.a) and the fact [1, 10] that, for  $\Omega \subset \mathbb{R}^2$  and for any  $1 < p < \infty$ ,

$$(3.7) \quad \|x\|_p \leq K \|x\|_1 ,$$

we use the Sobolev Imbedding Theorem [1, 10] to see that

$$\begin{aligned} \left| \int_0^t T_2 d\tau \right| &\leq \|u\|_{L^\infty(L^{2-\epsilon})} \int_0^t \|\nabla u\|_{L^{2+\epsilon_1}} \|\xi\|_{L^p} d\tau \\ &\leq K_2 h^{-\frac{\epsilon_1}{2+\epsilon_1}} K_1 \|\nabla u\|_{L^2(L^2)} \|\xi\|_{L^2(J_+, H^1)} \\ (3.8) \quad &\leq K h^{1-\epsilon - \frac{\epsilon_1}{2+\epsilon_1}} [\|\xi\|_{L^2(J_+, L^2)} + \|\nabla \xi\|_{L^2(J_+, L^2)}] \\ &\leq \frac{a}{8} (\|\nabla \xi\|_{L^2(J_+, L^2)}^2 + \|\xi\|_{L^2(J_+, L^2)}^2)^{2(1-\epsilon - \frac{\epsilon_1}{2+\epsilon_1})} + K h \end{aligned}$$

where  $\epsilon$ ,  $\epsilon_1$ , and  $p$  satisfy the relation

$$(3.9) \quad \frac{1}{2-\epsilon} + \frac{1}{2+\epsilon_1} + \frac{1}{p} = 1 .$$

We note that since  $\epsilon > 0$  from (2.5.c) can be arbitrarily small and arbitrarily large  $p$  satisfies (3.7),  $\epsilon_1$  can also be taken arbitrarily small and still satisfy (3.9). We next use (2.7.b-c) to obtain

$$(3.10) \quad \left| \int_0^t T_4 d\tau \right| \leq \frac{\alpha}{8} \sum_{j=1}^N \int_0^t |Q_j(\tau)| \xi^2(x_j, \tau) d\tau + K h^{2(1-\epsilon)}.$$

Finally we shall break  $T_3$  into pieces to consider as follows:

$$(3.11) \quad \begin{aligned} T_3 &= ((u - U) \cdot \nabla \xi, \xi) - ((u - U) \cdot \nabla \eta, \xi) + ((u - U) \cdot \nabla c, \xi) \\ &= T_5 + T_6 + T_7. \end{aligned}$$

Now we again use the Sobolev Imbedding Theorem with  $\epsilon_2 > 0$  arbitrarily small and  $p_2 > 0$  arbitrarily large satisfying

$$(3.12) \quad \frac{1}{2} + \frac{1}{2 + \epsilon_2} + \frac{1}{p_2} = 1$$

and apply (2.3.a), (2.13), and (3.7) to obtain

$$(3.13) \quad \begin{aligned} \left| \int_0^t T_5 d\tau \right| &\leq \|u - U\|_{L^\infty(L^2)} \int_0^t \|\nabla \xi\|_{L^2}^{2+\epsilon_2} \|\xi\|_{L^{p_2}} d\tau \\ &\leq K_5 h^{1-\epsilon} K_1 h^{-\frac{\epsilon_2}{2+\epsilon_2}} \|\nabla \xi\|_{L^2(J_t, L^2)} \|\xi\|_{L^2(J_t, H^1)} \\ &\leq K h^{1/2} \|\nabla \xi\|_{L^2(J_t, L^2)} [\|\xi\|_{L^2(J_t, L^2)} + \|\nabla \xi\|_{L^2(J_t, L^2)}] \\ &\leq \frac{\alpha}{8} [\|\nabla \xi\|_{L^2(J_t, L^2)}^2 + \|\xi\|_{L^2(J_t, L^2)}^2]. \end{aligned}$$

In (3.13) we have chosen  $\epsilon$  and  $\epsilon_2$  sufficiently small that  $\epsilon + \frac{\epsilon_2}{2+\epsilon_2} < \frac{1}{2}$ . In the same fashion, we use (2.3), (2.7a), and (2.13) to see that

$$\begin{aligned}
 | \int_0^t T_6 d\tau | &\leq \|u - U\|_{L^\infty(L^2)} \int_0^t \|\nabla n\| \|\xi\|_{L^\infty} d\tau \\
 (3.14) \quad &\leq K_5 h^{1-\epsilon} K_2 h^{1-\epsilon} K_1 h^{-1} \|\xi\|_{L^2(J_+, L^2)} \\
 &\leq \frac{\alpha}{8} \|\xi\|_{L^2(J_+, L^2)}^2 + K h^{2(1-2\epsilon)}.
 \end{aligned}$$

Now since (2.5.a) holds for  $\epsilon > 0$  arbitrarily small, an imbedding result similar to the one used in (3.7) can be applied to pick a  $p_3 = p_3(\epsilon) > 0$ , arbitrarily large, and satisfying

$$(3.15) \quad \|\nabla c\|_{L^2(L^{p_3})} \leq K_2$$

from (2.5.a). Using this  $p_3$ , we choose  $\epsilon_3 = \epsilon_3(p_3, \epsilon) > 0$ , arbitrarily small, to satisfy

$$(3.16) \quad \frac{1}{2} + \frac{1}{p_3} + \frac{1}{2 + \epsilon_3} = 1.$$

Then we see that, as before,

$$\begin{aligned}
 | \int_0^t T_7 d\tau | &\leq \|u - U\|_{L^\infty(L^2)} \int_0^t \|\nabla c\|_{L^{p_3}} \|\xi\|_{L^2}^{2+\epsilon_3} d\tau \\
 (3.17) \quad &\leq K_5 h^{1-\epsilon} \|\nabla c\|_{L^2(L^{p_3})} K_1 h^{-\frac{\epsilon_3}{2+\epsilon_3}} \|\xi\|_{L^2(J_+, L^2)} \\
 &\leq \frac{\alpha}{8} \|\xi\|_{L^2(J_+, L^2)}^2 + K h^{2(1-\epsilon - \frac{\epsilon_3}{2+\epsilon_3})}.
 \end{aligned}$$

We next combine the above estimates to see that for each  $t \in [0, T]$ , we have

$$(3.18) \quad \begin{aligned} \|\xi(t)\|^2 + \|\nabla \xi\|_{L^2(J, L^2)}^2 &+ \sum_{j=1}^N \int_0^t |Q_j(\tau)| |\xi^2(x_j, \tau)| d\tau \\ &\leq \|\xi\|_{L^2(J, L^2)}^2 + K h^{2(1-\hat{\epsilon})} \end{aligned}$$

where  $\hat{\epsilon}$  is defined as

$$(3.19) \quad \hat{\epsilon} = \max[2\epsilon, \epsilon + \frac{\epsilon_1}{2+\epsilon_1}, \epsilon + \frac{\epsilon_3}{2+\epsilon_3}]$$

and can be taken arbitrarily small. We can now apply Gronwall's lemma to (3.18) and use (2.7) and the triangle inequality to obtain the desired result (3.1).

We next consider Case II where  $\alpha_\ell > 0$  and  $\alpha_+ > 0$  model physical dispersion [12]. We obtain a reduced rate of convergence in this more complex case.

Theorem 3.2. Let  $(c, p)$  satisfy (1.1)-(1.2) and  $(C, P)$  satisfy (2.9), (2.10), and (2.14). There exist positive constants  $K_7 = K_7(\Omega, K_i; i=0, \dots, 5)$  and  $h_0$  such that, if  $h < h_0$ ,

$$(3.20) \quad \begin{aligned} \|c - C\|_{L^\infty(J, L^2)} + \|\nabla(c - C)\|_{L^2(J, L^2)} \\ + \left\{ \sum_{j=1}^N \int_0^T |Q_j(t)| |(c - c^2)(x_j, t)| dt \right\}^{1/2} \leq K_7 h^{1/2-\bar{\epsilon}}. \end{aligned}$$

( $\bar{\epsilon} = \bar{\epsilon}(\epsilon) > 0$  is defined in (3.35) below and can be taken arbitrarily small.)

Proof: Let  $\xi$  and  $n$  be as in the proof of Theorem 3.1. Subtracting (2.6) from (2.14) in this Case II and substituting  $\xi$  for  $x$  yields

$$\left( \phi \frac{\partial \xi}{\partial t}, \xi \right) + \sum_{i=1}^2 \sum_{j=1}^2 \left( D_{ij}(U) \frac{\partial}{\partial x_j} \xi, \frac{\partial}{\partial x_i} \xi \right) + (u \cdot \nabla \xi, \xi)$$

$$= (u \cdot \nabla \eta, \xi) + ((u - U) \cdot \nabla C, \xi)$$

$$(3.21) \quad N/2 \\ + \sum_{j=1}^{N/2} Q_j(t)(c - C)(x_j, t) \xi(x_j, t)$$

$$+ \sum_{i=1}^2 \sum_{j=1}^2 \left( D_{ij}(u) \frac{\partial}{\partial x_j} c - D_{ij}(U) \frac{\partial}{\partial x_j} \hat{c}, \frac{\partial}{\partial x_i} \xi \right)$$

$$\equiv T_8 + T_9 + T_{10} + T_{11}.$$

We again integrate (3.21) termwise on  $\tau$  in  $J_+ = [0, t]$  for  $t \in J$ . We obtain an analogue of (3.5) for the left-hand side of (3.21) where now  $\alpha$  depends upon the constant  $D_*$  assumed in (1.5). All the terms are then treated exactly as in the proof of Theorem 3.1 except for  $T_{11}$  which did not appear in Case I. We first split  $T_{11}$  up as follows:

$$(3.22) \quad T_{11} = \sum_{i=1}^2 \sum_{j=1}^2 \left( [D_{ij}(u) - D_{ij}(U)] \frac{\partial}{\partial x_j} c, \frac{\partial}{\partial x_i} \xi \right) \\ - \sum_{i=1}^2 \sum_{j=1}^2 \left( D_{ij}(U) \frac{\partial}{\partial x_j} \eta, \frac{\partial}{\partial x_i} \xi \right) \\ \equiv T_{12} + T_{13}.$$

We first note that

$$\begin{aligned}
 & \left| \int_0^t T_{13} d\tau \right| \leq 4 \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t (D_{ij}(U) \frac{\partial}{\partial x_j} n, \frac{\partial}{\partial x_i} n) d\tau \\
 (3.23) \quad & + \frac{1}{4} \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t (D_{ij}(U) \frac{\partial}{\partial x_j} \xi, \frac{\partial}{\partial x_i} \xi) d\tau \\
 & \equiv T_{14} + T_{15}.
 \end{aligned}$$

Clearly,  $T_{15}$  can be subtracted from the corresponding term on the left-hand side of (3.21). We can then split  $T_{14}$  as follows:

$$\begin{aligned}
 |T_{14}| & \leq 4 \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t ([D_{ij}(U) - D_{ij}(u)] \frac{\partial}{\partial x_j} n, \frac{\partial}{\partial x_i} n) d\tau \\
 (3.24) \quad & + 4 \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t (D_{ij}(u) \frac{\partial}{\partial x_j} n, \frac{\partial}{\partial x_i} n) d\tau \\
 & \equiv T_{16} + T_{17}.
 \end{aligned}$$

In order to bound  $\left| \int_0^t T_{12} d\tau + T_{16} \right|$ , we shall use (3.15) and an analogue for  $n$ . First  $c - c_1$  satisfies a bound of the form

$$(3.25) \quad \|\nabla(c - c_1)\|_{L^2(L^{p_3})} \leq K$$

where  $p_3$  is the same as in (3.15). Then since  $c_1 - \hat{c} \in M_h$ , we can use (2.3), (2.5), and (2.7) to see that, for  $\epsilon < \frac{2}{p_3}$ ,

$$\|\nabla(c_1 - \hat{c})\|_{L^2(L^{p_3})} \leq K_1 h^{\frac{2}{p_3}-1} \|\nabla(c_1 - \hat{c})\|_{L^2(L^2)}$$

$$(3.26) \quad \leq K_1 h^{\frac{2}{p_3}-1} K_2 h^{1-\epsilon} \|c\|_{L^2(H^{2-\epsilon})}$$

$\leq K$

Combining (3.25) and (3.26), we have

$$(3.27) \quad \|\nabla n\|_{L^2(L^{p_3})} \leq K.$$

We next note that by elementary but tedious computations one can show that  $D_{ij}(x, u)$  is Lipschitz in  $u$  with Lipschitz constant 3. Thus combining (3.15) and (3.27) and the Lipschitz behavior of  $D_{ij}$ , we can use (2.3.a) and (2.13) to obtain

$$(3.28) \quad \begin{aligned} & \left| \int_0^t T_{12} d\tau + T_{16} \right| \leq \|u - u\|_{L^\infty(L^2)} \int_0^t [\|\nabla n\|_{L^{p_3}} + \|\nabla c\|_{L^{p_3}}] \|\nabla \xi\|_{L^{2+\epsilon_3}} d\tau \\ & \leq K_5 h^{1-\epsilon} K K_1 h^{-\frac{\epsilon_3}{2+\epsilon_3}} \|\nabla \xi\|_{L^2(J_+, L^2)} \\ & \leq \frac{\alpha}{8} \|\nabla \xi\|_{L^2(J_+, L^2)}^2 + K h^{2(1-\epsilon-\frac{\epsilon_3}{2+\epsilon_3})}. \end{aligned}$$

We next use (1.7) and (2.12.a) to note that for each  $i, j = 1, 2$ ,

$$(3.29) \quad |D_{ij}(u)| \leq |u| \leq \sum_{k=1}^N K_8 \frac{1}{|x - x_k|} + |\alpha(x) \cdot \nabla \tilde{p}|$$

where  $\|v\|$  for a vector  $v$  is the standard Euclidean norm in  $\mathbb{R}^2$ . Using (2.3), (2.7), (2.8) and (3.29), we see that

$$\begin{aligned}
 |T_{17}| &\leq |T_{18}| + \|\nabla \tilde{p}\|_{L^\infty(L^4)} \int_0^t \|\nabla n\|_{L^4} \|\nabla n\| d\tau \\
 &\leq |T_{18}| + K h^{-\frac{1}{2}} \|\nabla n\|_{L^2(L^2)}^2 \\
 (3.30) \quad &\leq |T_{18}| + K h^{-\frac{1}{2}} h^{2(1-\epsilon)} \\
 &\leq |T_{18}| + K h^{2(\frac{3}{4}-\epsilon)}.
 \end{aligned}$$

We note that the second term on the right side of (3.30) does not give an optimal estimate for the  $a(x)\nabla \tilde{p}$  term from (3.29), but yields a better bound than we are at present able to obtain for  $T_{18}$ .  $T_{18}$  contains a term of size  $|x - x_j|^{-1}$  centered at each well. For simplicity, we will carefully estimate only one such term. Without loss of generality assume we have a well centered at the origin  $x = 0$ . We shall then split this term by considering the spatial integration over  $B_h^\beta$ , a disc of radius  $h^\beta$  centered at the origin, and its complement  $\Omega - B_h^\beta$ . We then see that

$$\begin{aligned}
 |T_{18}| &\leq N K \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t \int_{B_h^\beta} \frac{1}{r} \frac{\partial n}{\partial x_j} \frac{\partial n}{\partial x_i} dx d\tau \\
 (3.31) \quad &+ N K \sum_{i=1}^2 \sum_{j=1}^2 \int_0^t \int_{\Omega - B_h^\beta} \frac{1}{r} \frac{\partial n}{\partial x_j} \frac{\partial n}{\partial x_i} dx d\tau \\
 &\equiv T_{19} + T_{20}.
 \end{aligned}$$

We then obtain

$$(3.32) \quad \begin{aligned} |T_{20}| &\leq K h^{-\beta} \|v_n\|_{L^2(L^2)}^2 \\ &\leq K h^{-\beta} h^{2(1-\epsilon)} \leq K h^{2-2\epsilon-\beta}. \end{aligned}$$

Next, we let  $p_3$  and  $q$  satisfy

$$(3.33) \quad \frac{2}{p_3} + \frac{1}{q} = 1$$

and use (3.27) to obtain

$$(3.34) \quad \begin{aligned} |T_{19}| &\leq K \|v_n\|_{L^2(L^{p_3})}^2 \int_0^{2\pi} \left( \int_0^{h^\beta} r^{-q+1} dr \right)^{\frac{1}{q}} d\theta \\ &\leq K (h^\beta)^{\frac{-q+2}{q}} = K (h^\beta)^{1 - \frac{4}{p_3}} \\ &= K h^{\beta - \frac{4\beta}{p_3}} \end{aligned}$$

for  $p_3$  arbitrarily large. We then pick  $\beta$  to balance (3.32) and (3.34). With  $\beta = 1$  and

$$(3.35) \quad \bar{\epsilon} = \max \left[ \epsilon, \frac{2}{p_3} \right],$$

we see that

$$(3.36) \quad |T_{18}| \leq K h^{2(\frac{1}{2} + \bar{\epsilon})}.$$

Combining the above estimates and corresponding bounds from the proof of Theorem 3.1, we obtain, for each  $t \in [0, T]$ .

$$(3.37) \quad \|\xi(t)\|^2 + \|\nabla \xi\|_{L^2(J_t, L^2)}^2 + \sum_{j=1}^N \int_0^t |Q_j(\tau)| \xi^2(x_j, \tau) d\tau$$

$$\leq \|\xi\|_{L^2(J_t, L^2)}^2 + K h^{2(\frac{1}{2} - \bar{\varepsilon})}$$

where  $\bar{\varepsilon}$  is given by (3.36) and can be taken arbitrarily small. Then applying Gronwall's lemma to (3.37) we use (2.7) and the triangle inequality to obtain the desired result, (3.20).

References

1. R. A. Adams, Sobolev Spaces, Academic Press, New York, 1975.
2. J. H. Bramble and S. R. Hilbert, "Bounds for a class of linear functionals with applications to Hermite interpolation", Numer. Math. 16, pp. 362-369.
3. J. Douglas, Jr., "The numerical solution of miscible displacement in porous media", Computational Methods in Nonlinear Mechanics, (J. T. Oden, ed.), North Holland, New York, 1980.
4. J. Douglas, Jr., T. Dupont, and R. E. Ewing, "Incomplete iteration for time-stepping a Galerkin method for a quasilinear parabolic problem", SIAM J. Numer. Anal. 16 (1979), pp. 503-522.
5. J. Douglas, Jr., R. E. Ewing, and M. F. Wheeler, "Mixed methods for miscible displacement problems in porous media", (to appear).
6. J. Douglas, Jr., M. F. Wheeler, B. L. Darlow, and R. P. Kendall, "Self-adaptive finite element simulation of miscible displacement in porous media", SIAM J. Sci. Stat. Comp. (to appear).
7. R. E. Ewing and M. F. Wheeler, "Galerkin methods for miscible displacement problems in porous media", SIAM J. Numer. Anal. 17 (1980), pp. 351-365.
8. R. E. Ewing and T. F. Russell, "Efficient time-stepping procedures for miscible displacement problems in porous media", SIAM J. Numer. Anal. (to appear).
9. P. Grisvard (private communication).
10. J. L. Lions and E. Magenes, Non-homogeneous Boundary Value Problems and Applications, Vol. I, Springer-Verlag, New York, 1972.

11. D. W. Peaceman, Fundamentals of Numerical Reservoir Simulation, Elsevier Publishing Company, 1977.
12. D. W. Peaceman, "Improved treatment of dispersion in numerical calculation of multidimensional miscible displacement", Soc. Pet. Eng. J. (1966), pp. 213-216.
13. T. F. Russell, "An incompletely iterated characteristic finite element method for a miscible displacement problem", Ph.D. Thesis, University of Chicago, Chicago, 1980.
14. P. H. Sammon (private communication).
15. A. Settari, H. S. Price, and T. Dupont, "Development and application of variational methods for simulation of miscible displacement in porous media", Soc. Pet. Eng. J. (June 1977), pp. 228-246.
16. M. F. Wheeler, "A priori  $L^2$ -error estimates for Galerkin approximations to parabolic partial differential equations", SIAM J. Numer. Anal. 10 (1973), pp. 723-759.
17. M. F. Wheeler and B. L. Darlow, "Interior penalty Galerkin methods for miscible displacement problems in porous media", Computational Methods in Non-Linear Mechanics, J. T. Oden, editor, North Holland Publishing Company, New York (1980), pp. 485-506.

Tracking of interfaces in fluid flow: Accurate methods for  
piecewise smooth problems

James Glimm

ABSTRACT

A survey of hyperbolic conservation laws is presented, with an emphasis on issues raised by a front tracking code developed by the author, Eli Isaacson, D. Marchesin and O. McBryan. The organization of the code is described and results of the calculations are summarized. The aim of the code is to provide a general and flexible method for obtaining accurate solutions to problems which are piecewise smooth.

I INTRODUCTION

A number of problems of fundamental importance to science and technology involve an interdisciplinary mix of fluid dynamics, physics and/or chemistry and computer modelling. The fluids may be either liquids or gases. The flow can be turbulent or laminar. It can have boundary layers, detached boundary layers, or internal fluid and material discontinuities (shock and contact waves). In addition to exhibiting such pure fluid phenomena, fluids of interest may do something

besides just flowing. The fluid constituents may react chemically (e.g. burn), they may change phase, precipitate as solids, or become adsorbed at the active site of a catalyst in a reactor bed.

Because computer modelling is an essential part of the problems I will discuss, it is necessary to make two comments on  $\Delta x$ , as an introduction to the methods to be proposed. First,  $\Delta x$  does not go to zero. Second,  $\Delta x$  does not vary greatly. Working with a \$100K mini-computer, it is fairly routine to solve time dependent problems containing a two dimensional elliptic equation on a  $30 \times 30$  mesh. With a large computer, the cost would be several million dollars, and a typical grid, might be  $80 \times 80$  or even  $150 \times 150$ , depending on the computer and the importance of the problem. For three dimensions, the grids are correspondingly coarser. These grid sizes are sometimes adequate to resolve the principal hydrodynamics waves in a complex problem, especially for a two dimensional calculation. They are almost never sufficient to resolve secondary waves, nor are they sufficient to resolve physical or chemical processes which occur on length scales much smaller than those of the principal hydrodynamics waves, unless some special adaptive strategy is employed.

A second degression is required to explain some consequences of the basic scientific phenomena to be modelled. Many problems have the form of a hyperbolic conservation law (mass, momentum, chemical species, ...) with an elliptic term added (heat conduction, mass diffusion, viscosity, ...). These problems are parabolic, but depending on the relative size of the parameters, they may be regarded as approximately hyperbolic or elliptic. In fact for a problem with several degrees of freedom, some of the degrees of freedom may be approximately elliptic, while others may be approximately hyperbolic.

Let us consider a hyperbolic degree of freedom. This means that the associated diffusion length is significantly smaller than one mesh spacing; since the mesh spacing is not likely to change by as much as a factor of eight from a

small scale to a large scale calculation, this property is independent of mesh spacing over a practical range of mesh choices. As a temporary approximation, we set the diffusion length to zero. In some cases (the stable regime), the limit of zero diffusion length is continuous. Still the numerical implementation of zero diffusion requires special methods. In fact the most common numerical methods, even when applied to problems which are stable physically as far as can be determined by experiment and by linear stability analysis of special solutions, exhibit instabilities in the zero diffusion limit. These instabilities are numerical and not physical; they are properties of the discrete approximation and the solution algorithm, and not of the continuum equations nor the physics which they model. To avoid these instabilities, a minimum diffusion length of at least two and perhaps up to four mesh spaces is required. This diffusion length is a numerical artifact. It is perhaps the largest consistent error in numerical hydrodynamics.

Remaining in the (physically) stable regime, let us now consider a nonzero diffusion length. In this case the discontinuity wave is replaced by a smooth transition in a narrow region with steep gradients. Actually the situation as seen by the experimental physicist or chemist is often more complicated. The equations which we originally set out to solve are a projection onto a small number of degrees of freedom of a very large system. The internal structure of a single discontinuity wave may be a series of sharply defined waves of the larger system, moving with a common velocity. Flame fronts are commonly of this nature. When the internal structure of a hyperbolic wave consists of subwaves of extra degrees of freedom, then the correction to include this phenomena can be performed within a framework of sharply resolved discontinuities. However, the diffusion may also be real, i.e. the experimentally correct source of internal structure for the wave, as in the case of a drop of dye in clear water.

Finally we discuss the unstable regime. In this case, the use of a zero diffusion length, if taken literally, would give an incorrect solution. In this case, the subgrid phenomena, which occur on length scales too small to be computed directly but which still govern the physical stability of the flow and thus affect the large scale hydrodynamic waves, must be somehow retained or replaced by some effective numerical equivalent. In addition to parabolic terms with an associated diffusion length, subgrid effects may include surface tension and heterogeneity (random media).

There is a computational strategy which allows zero numerical diffusion. It is to track the waves. Singularities (e.g. discontinuities) of a solution of a hyperbolic equation propagate along characteristics. The characteristics are solutions of an ordinary differential equation determined by the wave speeds, and for a nonlinear equation, by the solution itself. It is possible to use a purely characteristic approach, in which all waves, singular or smooth, are propagated along characteristics. This is known as the method of lines. The method of tracking is a hybrid, which uses the characteristic propagation for certain waves (the "tracked waves") and a finite difference grid for the other waves. Specializing to a two dimensional problem we then have a fixed two dimensional rectangular or curvilinear grid for the untracked (smooth) waves and a moving one dimensional curvilinear grid which locates the position of the tracked wave. This method can be viewed as a variant of the adaptive grid and mesh refinement approaches. In fact when the internal structure of the tracked wave is parabolic and governed by a mixing length (rather than containing several waves from new hyperbolic degrees of freedom), and when this internal structure is required as part of the solution of the problem, then the methods of moving grid local mesh refinement and of tracking overlap.

In summary, the computational methods which employ the known analytic and qualitative properties of the solution have the promise of achieving increased accuracy, speed and especially resolution.

## II THEORY

We focus on elementary waves and their interactions. The Riemann problem is basic, and we explain in what ways a deeper understanding is required.

### Equations

The equations of (chemically reacting, ...) fluid dynamics are basically conservation laws. The simplest is the continuity equation

$$\frac{u}{t} + \vec{\nabla} \cdot (\vec{v} u) = 0 \quad (2.1)$$

for a quantity (concentration, mass, ...)  $u$  carried passively in a fluid with a velocity  $\vec{v}$ . Writing

$$(\frac{u}{t}, \vec{v}) = D \quad (2.2)$$

as the space time gradient, we see that (2.1) states that the vector

$$(u, \vec{v} u) \quad (2.3)$$

has zero space time divergence:

$$D(u, \vec{v} u) = 0 .$$

Thus the vector  $(u, \vec{v} u)$  is a conserved quantity. By the divergence theorem, the integral of the outward normal component of this vector over the boundary  $\partial\Omega$  of any region vanishes,

$$\oint_{\partial\Omega} (u, \vec{v} u) \cdot \hat{n} d\sigma = 0 . \quad (2.5)$$

In particular we choose

$$\Omega = [t_1, t_2] \times \hat{\Omega} \quad (2.6)$$

to be a cylinder of height  $t_2 - t_1$  and base  $\hat{\Omega}$ . Then

$$\int_{\hat{\Omega}} u(\vec{x}, t_1) d\vec{x} - \int_{\hat{\Omega}} u(\vec{x}, t_2) d\vec{x} = \int_{t_1}^{t_2} \int_{\partial\hat{\Omega}} u \vec{v} \cdot \hat{n} d\sigma$$

Note that the left hand side is the quantity of  $u$  at time  $t_1$  minus that at time  $t_2$ , i.e. the change in the amount of  $u$ , while the right hand side is the flux of  $u$  across the boundary  $\partial\hat{\Omega}$  of  $\hat{\Omega}$ . In other words,

$$\text{change in } u = \text{flux across boundary.} \quad (2.8)$$

Of course the reasoning can be reversed. Formula (2.8) can be taken as fundamental, i.e. the definition of a conserved quantity, and (2.1) can be derived from it by considering all possible  $\Omega$ 's and taking a limit as  $\Omega$  becomes infinitesimal.

More generally we consider quantities which need not be carried passively by the fluid flow. For example, pressure, or acoustical waves in a gas move (by definition) with the speed of sound, while the gas molecules move with the gas (wind) velocity. In general, the velocity of a chemical reaction wave is different from the velocity of the molecules because in the reaction the molecules are changing species, and thus distinct molecules are located at the wave reaction front at distinct times. Thus we introduce a general flux function

$$\vec{f} = \vec{f}(u) \text{ or } \vec{f} = \vec{f}(u, t, \vec{x})$$

Note that  $\vec{x} \in R^d$  is a vector taking values in the geometrical - physical - space, while  $u$  is also a vector, but takes its values in the state space  $R^n$  which defines the degrees of freedom of the problem (momentum, mass, energy, ...). The values of  $\vec{f}$  lie in  $R^d \times R^n$ . Reasoning as above from the definition of a conservation law, we are lead to the conservation law equation

$$u_t + \vec{\nabla} \cdot \vec{f}(u) = 0 . \quad (2.9)$$

An external source  $g = g(u, t, \vec{x})$ , if any, replaces zero on the right side of (2.9). In (2.9) with source  $g = 0$ ,  $u$  is a conserved quantity and  $\vec{f}$  is its flux function. In other words  $\vec{f} \cdot \vec{n}$  is the rate of flow of  $u$  across a unit surface element with unit normal  $\vec{n}$ .

A preliminary classification of the equation (2.9) falls back on the linear theory, and so we introduce the Jacobean

$$\vec{A} = \partial \vec{f} / \partial u = (\partial \vec{f}_i / \partial u_j) , \quad (2.10)$$

$1 \leq i, j \leq n$  and

$$\vec{A} \cdot \vec{n} = \partial(\vec{f} \cdot \vec{n}) / \partial u . \quad (2.11)$$

Then

$$w_t + \vec{A} \cdot \vec{\nabla} w + (\vec{f} \cdot \vec{A}(u_0)) w = 0 \quad (2.12)$$

is the linearization of (2.9) about the solution  $u$ . For plane waves moving in the direction  $\vec{n}$ , the equation specializes to

$$w_t + (\vec{A} \cdot \vec{n}) (\vec{n} \cdot \vec{\nabla}) w = 0 . \quad (2.13)$$

Note that  $\vec{A} \cdot \vec{n}$  is an  $n \times n$  matrix.

Following standard linear terminology, we say that (2.9) is hyperbolic if all the eigenvalues  $\lambda_i = \lambda_i(u, \vec{n})$  are real and it is strictly hyperbolic if the  $\lambda_i$  are real and distinct. In general  $A$  is not symmetric. Let  $e_i^l$  and  $e_i^r$  denote the left and right eigenvectors corresponding to the eigenvalue  $\lambda_i$ , so that

$$e_i^l \vec{A} \cdot \vec{n} = \lambda_i e_i^l , \vec{A} \cdot \vec{n} e_i^r = \lambda_i e_i^r . \quad (2.14)$$

Then the  $e_i^l$  and  $e_i^r$  form a biorthogonal family:

$$\langle e_i^l, e_j^r \rangle = 0 \text{ for } \lambda_i \neq \lambda_j . \quad (2.15)$$

The equation (2.9) is linear if and only if  $\vec{A} \cdot \vec{n}$  is independent of  $u$ . Conversely, we say that (2.9) is strictly nonlinear if

$$\nabla_{e_i^r} \lambda_i (u, \vec{n}) \neq 0 \quad (2.16)$$

for all  $i$ . If it is strictly hyperbolic, so that wave speed crossings do not occur and the  $i^{th}$  mode is globally defined, then we say that the  $j^{th}$  mode is strictly nonlinear if (2.16) holds for  $i = j$ . The importance of this concept will emerge later, but for now, we specialize to  $n = 1$ , a scalar equation. Then

$$\lambda_i = \vec{A} \cdot \vec{n} = (\partial \vec{f} / \partial u) \cdot \vec{n} \quad (2.17)$$

is real valued,  $e_i^r = 1$ , and

$$\nabla_{e_i^r} \lambda_i = (\partial^2 \vec{f} / \partial u^2) \cdot \vec{n} . \quad (2.18)$$

In other words, strictly nonlinear means that  $\vec{f} \cdot \vec{n}$  is either concave or convex as a function of  $u$  in the scalar case.

There are important examples which are strictly hyper-

bolic, others which are hyperbolic but not strictly hyperbolic. The same applies to the strictly nonlinear concept. Probably there are very few properties of  $A$  which are universal to all examples, but it is an open question to determine properties of  $A$  or  $f$  which apply to various class of natural examples and which allow the interaction of elementary waves (or the Riemann problem) to be understood in the large. It is this question to which we now turn.

#### Elementary waves.

The easiest way to appreciate the meaning of elementary waves is to study the Riemann problem. The Riemann problem is the initial value problem for one space dimension and for data which is constant except for a single jump discontinuity:

$$u_0 = u(t=0, x) = \begin{cases} u_{\text{left}}, & x < 0 \\ u_{\text{right}}, & x > 0 \end{cases} \quad (2.19)$$

The solution

$$u(t, x) = u(1, x/t) \quad (2.20)$$

is constant on rays and reduces to a function of one variable:  $\xi = x/t$ . To see this observe that the equation (2.9) is invariant under the scale transformation

$$x \rightarrow ax, t \rightarrow at \quad (2.21)$$

Assuming uniqueness (which has been proved in a large enough class of problems to apply here [DiPerna, 1979]), and observing that the data (2.19) is also scale invariant, then so is the solution scale invariant. Thus (2.20) holds. For an  $n$ -dimensional state space the solution (in the simplest case; exceptions will be considered later) consists of  $n+1$  wedges in which  $u$  is constant. Between adjacent wedges in which  $u$  is constant, the allowed change in  $u$  is either a single jump discontinuity or smooth change in an intervening wedge of a type to be presecribed below. In either case, the variation of  $u$  between adjacent wedges of constancy is an elementary wave. The  $i^{\text{th}}$  wave involves variation in a single mode, or

eigendirection  $e_i^r(u)$ , and propagates with a speed  $\lambda_i(u) \approx \xi = x/t$  at least for infinitesimal waves, as we shall see. In summary, each elementary wave

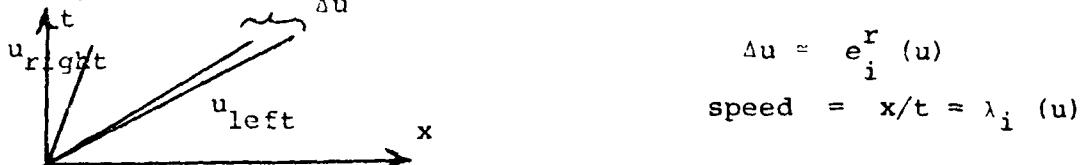


Figure 2.1 Elementary wave involves variation in a single mode, and propagates with a speed characteristic of that mode. The spreading waves are called rarefaction waves; the jump discontinuous waves are either shock waves or contact discontinuities, as we now explain.

Because we are contemplating discontinuous solutions, we are necessarily considering weak solutions. The concept of weak solution can be formulated in three equivalent ways. First, it can be required that the original conservation and flux relations which defined the conservation law be satisfied across discontinuities. Second, the integral form of the conservation law,

$$\int (\phi_t u + \phi_x f(x)) dt dx + \int \phi(0, x) u(0, x) dx = 0 \quad (2.22)$$

for all smooth  $\phi$  with compact support, defines a weak solution. Third, if the derivatives in (2.9) are taken in the sense of Schwartz distributions, then again a weak solution is defined. Now consider a curve  $x = x(t)$  in space-time moving with speed  $s = \dot{x}$ , and suppose that a solution  $u$  of (2.9) is discontinuous across  $x(t)$ . Apply the space-time divergence theorem to the vector field  $u, f(u)$  in a small strip  $\Omega$  around the curve  $x(t)$ .

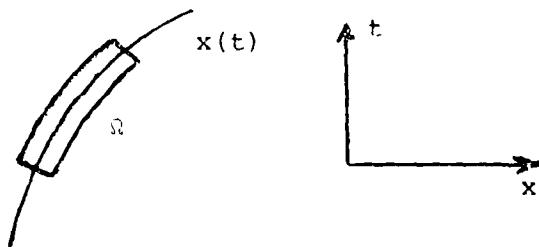


Figure 2.2. The Rankine Hugoniot relations derived from the space-time divergence theorem.

Note that  $(1, s)$  is tangent to the curve and  $(-s, 1)$  is normal to the curve. By the weak formulation of the differential equation, the divergence of  $u, f(u)$  vanishes when integrated over  $\Omega$  and so by the divergence theorem,

$$\int_{\partial\Omega} (u, f(u)) \cdot n \, ds = 0.$$

If we let  $[w]$  denote the jump in a quantity  $w$  across the curve  $x(t)$ , then by shrinking  $\Omega$ , we conclude that

$$([u], [f]) \cdot (-s, 1) = 0$$

identically along the curve. Thus

$$s[u] = [f], \quad (2.23)$$

which is known as the Rankine-Hugoniot relation.

If we specialize to an infinitesimal wave, then the above identity becomes

$$s \, du = df.$$

Recalling that  $A = df/du$ , we rewrite this as

$$s \, du = A \, du \quad (2.24)$$

So that we identify  $s$  with an eigenvalue  $\lambda_i$  of  $A$  and  $du$  with an infinitesimal variation in the direction of the eigenvector  $e_i^T(u)$ . Motivated by the formula (2.24), it is easy to construct the rarefaction waves. The construction begins in the state space. Let  $w_{left}$  be the value of  $u$  in the constancy wedge to the left of the  $i^{th}$  wave.

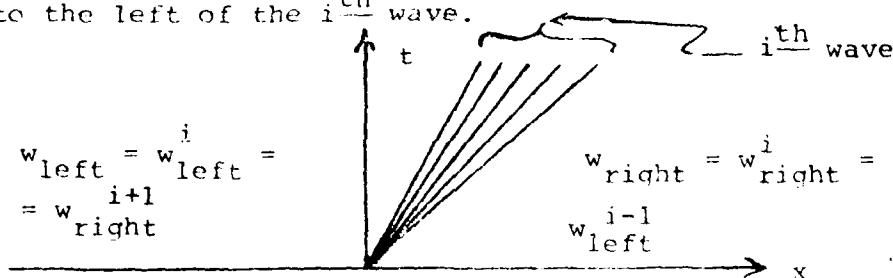


Figure 2.3 An elementary rarefaction wave.

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

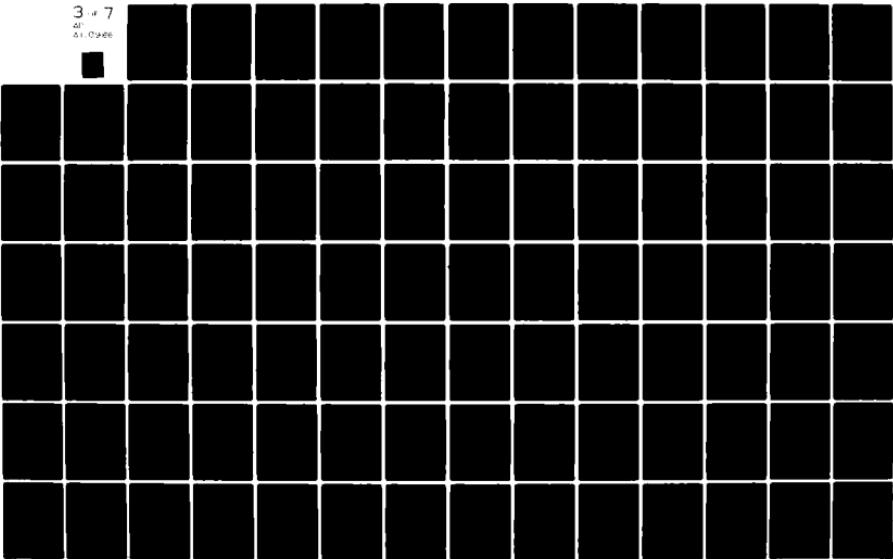
F/G 12/1

UNCLASSIFIED

AFOSR-TR-82-0047

NL

3 x 7  
AF  
AD-966



In the state space, we solve the ordinary differential equation

$$\dot{u} = e_i^r(u)$$

with initial point  $u_0 = w_{\text{left}}^i$ .

The equation is autonomous and so the integral curve does not depend on the point along it chosen as initial point. Assuming strict nonlinearity, the wave speed  $\lambda_i(u)$  is strictly increasing as we move in one direction along the integral curve. Now identifying  $\lambda_i = \xi = x/t$  in (2.29) defines the solution  $u$  in the region of the  $i^{\text{th}}$   $w_{\text{left}}^i$  and  $w_{\text{right}}^i$ . Note that in moving from left to right, we are constrained to move along the integral curve in the direction of increasing  $\lambda_j(u)$ .

The general strategy for solving the Riemann problem is to use the elementary wave strength as a parameter in passing from  $w_{\text{right}}^{i+1} = w_{\text{left}}^i$  to  $w_{\text{right}}^i$ . After  $n$  steps, we have a solution depending on  $n$  parameters, joining left and right states. The rarefaction waves allow only one sided variation of the parameters, but the shock waves, defined by (2.23) will provide variation in the opposite direction. Then an application of the implicit function theorem will show that an arbitrary state  $u_{\text{right}}$  satisfying

$$|u_{\text{left}} - u_{\text{right}}| < \epsilon$$

can be joined to  $u_{\text{left}}$  with this  $n$  parameter family of elementary waves, thereby solving the Riemann problem in the small. The corresponding problem in the large requires some hypothesis on  $\Lambda$ , and has been solved only in a few special cases.

An example: Burgers' equation. The simplest example of a conservation law is the scalar equation

$$u_t + (\frac{1}{2}u^2)_x = 0 , \quad (2.25)$$

known as Burgers' equation. Here the matrix  $A = \partial f / \partial u$  is a real number,

$$A = \frac{\partial (\frac{1}{2}u^2)}{\partial u} = u = \lambda_1 \quad (2.26)$$

Because  $\frac{\partial \lambda_1(u)}{\partial u} = 1 \neq 0$ , the equation is strictly non-linear as well as being strictly hyperbolic. It is elementary to construct shock and rarefaction waves to solve the Riemann problem for Burgers' equation. If  $u_{\text{left}} < u_{\text{right}}$ , then the solution is a rarefaction wave, defined as follows:

$$u(t, x) = u(1, x/t) = \begin{cases} u_{\text{left}} & \text{if } x/t \leq u_{\text{left}} \\ x/t & \text{if } u_{\text{left}} \leq x/t \leq u_{\text{right}} \\ u_{\text{right}} & \text{if } u_{\text{right}} \leq x/t . \end{cases} \quad (2.27)$$

If  $u_{\text{left}} > u_{\text{right}}$ , then the solution is a shock wave. By (2.23), we have

$$[u] = u_{\text{right}} - u_{\text{left}}$$

$$[f] = \frac{1}{2}(u_{\text{right}}^2 - u_{\text{left}}^2) = \frac{1}{2}[u] (u_{\text{right}} + u_{\text{left}})$$

and so

$$s = [f]/[u] = \frac{1}{2}(u_{\text{right}} + u_{\text{left}}) .$$

The value of  $s$  uniquely determines the solution. If we draw the characteristic curves  $\dot{x} = \lambda$  in the regions where  $u$  is constant, then we obtain the pictures below.

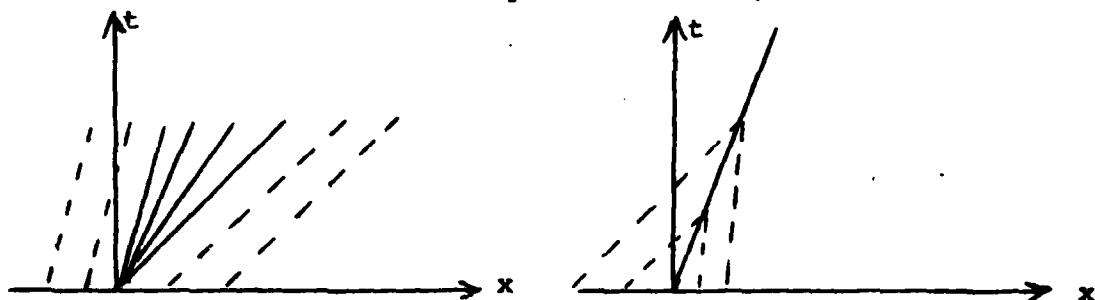


Figure 2.4a,b. Rarefaction and shock waves for Burger's equation. Dashed lines are characteristics. From this picture, we see that a shock wave is like a black hole: it absorbs information propagating along characteristics, and this information is then lost in the solution.

Entropy.

The equation (2.9) is invariant under space-time reversal,

$$x \rightarrow -x, t \rightarrow -t .$$

This reversal interchanges shock and rarefaction wave data, but it does not interchange shock and rarefaction wave solutions. In fact it maps a shock wave onto a solution which could be called a "rarefaction shock wave", and which we want to exclude. Since the rarefaction shock wave is a weak solution of the equation (2.9), a new condition is required to supplement (2.9). This condition is the entropy condition. The characteristics for the rarefaction shock are shown below;

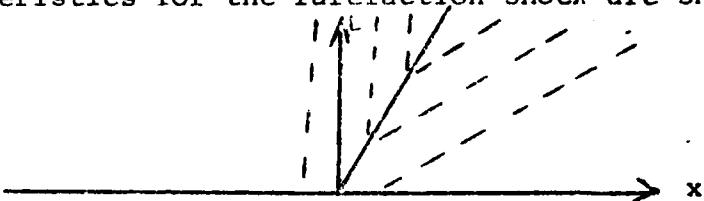


Figure 2.5 A (nonphysical) rarefaction shock. they are the  $x-t$  reversal of the shock characteristics shown above. One form of the entropy condition excludes discontinuities (shocks) of the  $i^{\text{th}}$  family in which the forward  $i^{\text{th}}$  family characteristics leave the shock from either side. Another form of the entropy condition states that there is no path in state space consisting of a sequence of elementary waves, joining  $u_{\text{left}}$  to  $u_{\text{right}}$  and with increasing wave speeds as the path is traversed from  $u_{\text{left}}$  to  $u_{\text{right}}$ . In other words the jump is indecomposable. A third form of the condition states that the solution should be the limit, as  $\epsilon \rightarrow 0$ , of solutions of the parabolic equation

$$u_t + f(u)_x = \epsilon u_{xx} .$$

In general, the rarefaction shocks are unstable and are excluded by perturbing the initial data or the equation. For general equations, considered in the large, it is not known which forms of the entropy condition will be correct.

Contact Discontinuities.

Linear waves and contact discontinuities do not arise in Burgers' equation, so we consider the linear equation

$$u_t + u_x = 0$$

with  $f = u$  and  $A = \partial f / \partial u = \lambda_1 = 1$ . The equation (2.24) shows that discontinuities can arise in the case  $[u] = [f]$  and  $s = \lambda = 1$ . Then the characteristics run parallel to the discontinuity curve, and neither enter it (as in a shock) nor leave (as in the excluded rarefaction shock). See below.

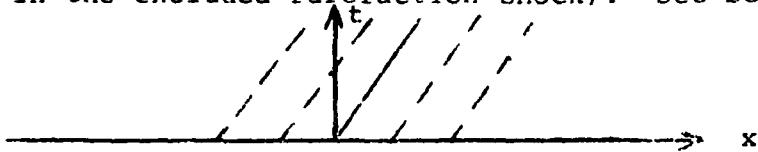


Figure 2.6 A contact discontinuity

The general definition of a contact discontinuity is a jump discontinuity satisfying (2.24) for which  $s = \lambda_i (u_{\text{left}}) = \lambda_i (u_{\text{right}})$ , as in the figure above.

#### The scalar equation: The general theory.

For the scalar equation, the Riemann problem can be solved for arbitrary  $u_{\text{left}}$ ,  $u_{\text{right}}$  without further hypothesis. For the strictly nonlinear case,  $f$  is convex or concave, and a slight extension of the case of Burgers' equation covers the situation. In the general case ( $f$  neither convex nor concave) we join  $u_{\text{left}}$  to  $u_{\text{right}}$  by a sequence of elementary waves.

In the  $u$ ,  $f(u)$  plane, we recognize a shock wave as a chord, joining two points on the graph of  $f$ . The speed of the shock wave is the slope of the chord. Also a rarefaction wave is a portion of the graph of  $f$ , and the local wave speed within the rarefaction wave is  $\lambda = f'(u)$ . Thus a sequence of elementary waves is just a sequence of chords and segments of the graph. This sequence must join  $u_{\text{left}}$ ,  $f(u_{\text{left}})$  to  $u_{\text{right}}$ ,  $f(u_{\text{right}})$ , subject to two constraints: the wave speeds must increase when moving from left to right and the entropy condition means that the elementary wave sequence forms a concave set in the  $u, f$  plane for  $u_{\text{left}} < u_{\text{right}}$  and a convex set for  $u_{\text{left}} > u_{\text{right}}$ . The entropy condition, which forbids waves which can be subdivided, forces the concave (or convex) set to be the concave (or convex) envelope of the graph of  $f$ , between  $u_{\text{left}}$  and  $u_{\text{right}}$ . The minimally complicated extension of this solution to the case of  $n \times n$  systems follows: Without

assuming strict nonlinearity (i.e. the general nonconvex, non-cave case), the single elementary wave of the  $i^{\text{th}}$  family in the Riemann problem solution may now be replaced by a sequence of elementary waves of the  $i^{\text{th}}$  family, alternately rarefaction waves and contact discontinuities, except for the outermost waves, which may be shocks when viewed from this outer direction. See below.

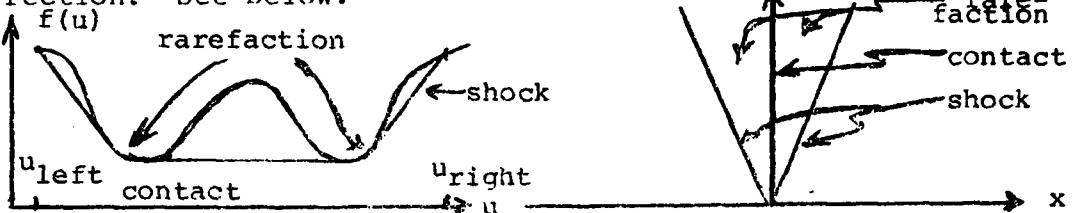


Figure 2.7 Elementary waves in the nonconvex case  
Under further hypothesis, this class of elementary wave solutions is adequate for strictly hyperbolic  $2 \times 2$  systems. The hypothesis may include most cases of interest for small data, but it is not known (and possibly false) that they will include most cases of interest for arbitrary left and right hand states. See [Wendroff, 1972, DaFermos, 1973, Liu, 1974]. The situation for  $n \times n$  systems,  $n > 2$ , is understood for small data [Liu, 1975].

#### The Riemann problem for small data.

For a general  $n \times n$  system, the Riemann problem can be solved for  $u_{\text{left}} = u_{\text{right}}$ , using the implicit function theorem.

#### Theorem.

Assume that the  $n \times n$  system of conservation laws (2.9) is strictly hyperbolic and strictly nonlinear. Let a left state  $u_{\text{left}}$  be given. Then for any right state  $u_{\text{right}}$  sufficiently close to  $u_{\text{left}}$ , the Riemann problem is solvable and the solution contains  $n$  elementary waves, each of which is either a rarefaction wave or a shock wave. The solution satisfies the entropy condition (no forward characteristics of the  $i^{\text{th}}$  family leave an  $i$ -family shock wave), and is the unique such solution in this class.

Proof [Lax, 1957].

The Riemann problem for arbitrary data.

There are only a few special systems for which a complete solution is given analytically. These include: Isentropic (2x2) gas dynamics [Codunov, 1959], Elasticity in Lagrange coordinates [Wendroff, 1972], and polymer injection in tertiary oil recovery [Eli Isaacson, 1981]. The polytropic (3x3) gas dynamics can be reduced analytically to a functional equation in one dimension which is easy to solve numerically [Courant-Friedrichs, 1948]. Other equations of state sufficiently similar to a polytropic gas are also allowed. In most other cases, one finds special solutions (i.e. solutions for special values of  $u_{\text{left}}$ ,  $u_{\text{right}}$ ) but no systematic analysis, either qualitative or numerical, of the general Riemann problem.

The polymer problem mentioned above and related earlier work [Keyfitz and Kranzer, 1980] concern systems which are hyperbolic but not strictly hyperbolic. In the polymer problem, the wave speeds cross, and coincide along a curve (the transition curve) in state space. When the solution to a Riemann problem crosses such a transition curve, an extra family of elementary waves may be required. Thus with a single crossing and a 2x2 system, three elementary wave families may be required. In the nonconvex case, each family may consist of rarefaction waves with imbedded contact discontinuities and one sided shocks at the outer edges, as in the case of a nonconvex scalar equation.

Solutions in the large for arbitrary data, d = 1.

For unrestricted (bounded variation) data, solutions in the large are known for single equations and for two special 2x2 systems: isothermal gas dynamics [Nishida, 1968] and polymer oil recovery [Temple, 1981]. For data with small oscillation, but otherwise unrestricted, there is a satisfactory general theory beginning with the papers [Glimm, 1965] and [Glimm and Lax, 1970]. Uniqueness [DiPerna, 1979], regularity [DiPerna, 1975] and large time asymptotics [Liu, 1981] are under control, although some aspects of the uniqueness question remain open. The use of an equidistributed sequence

as a sampling method in this construction was justified by [Liu, 1977]; the proof involves tracing of wave packets through the approximate solution and should be useful for other purposes, for example the problem of continuity in the initial data (which is open).

#### Interaction of waves in higher dimensions.

The general computational program outlined in the introduction is to incorporate analytic information concerning wave structure and wave propagation into the computational algorithm where possible. To do this, positions of certain "tracked" waves are stored and dynamically updated at each time step. The stored wave position defines locally a curvilinear coordinate system. In this coordinate system, the wave motion is essentially one dimensional and governed by a Riemann problem. Moreover, continuum waves contained in the smooth part of the solution can be resolved in this local coordinate system into normal and tangential components. The normal components interact with the tracked wave via a Riemann problem, while the tangential components move independently of the tracked wave.

The interaction of tracked waves, however, is intrinsically higher dimensional. If the interacting waves are not parallel, but meet obliquely, then the resulting wave configuration is not solved by a one dimensional Riemann problem. In fact, the interaction of obliquely intersecting waves in higher dimensions is the higher dimensional analog of the Riemann problem. It has been studied only in some special cases [Courant and Friedrichs, 1948].

In summary, we see that the theory of the Riemann problem in one and higher dimensions needs considerable development. The theory of general solutions in one space dimension is satisfactory but not complete.

#### III Computation.

Here we describe the front tracking [Glimm, Eli Isaacson, Marchesin and McBryan, 1981] and mesh alignment [McBryan, 1980] algorithms down to some intermediate level of detail.

The general scientific perspective on which they are based was explained in the introduction. We consider a specific problem: an oil reservoir undergoing water flood, modelled by Darcy's law and the Buckley-Leverett equation. The basic equations derive from conservation of mass of water and of oil. The equations can be added, and assuming incompressibility, the sum of the two conservations has no time derivative. This resulting equation is elliptic; it determines the pressure and fluid velocities (Darcy's law) from source terms, the relative oil/water saturations and viscosities and rock properties (absolute and relative permeabilities and porosity). It is

$$\vec{v}(x,y,t) = -k(s) \vec{\nabla} p \quad (3.1)$$

$$\vec{\nabla} \cdot \vec{v} = \text{source terms} . \quad (3.2)$$

Here  $s = s(x,y,t) \in [0,1]$  is the relative saturation of water in the porous media and  $k = k(s)$  (or  $= k(s,x,y)$ ) is an experimentally or phenomenologically determined function. The saturation, or Buckley-Leverett equation is a scalar hyperbolic conservation law

$$s_t + \vec{\nabla} \cdot (\vec{v} f(s)) = \text{source terms} . \quad (3.3)$$

Because (3.3) contains the velocity  $\vec{v}$ , the elliptic and hyperbolic equations are coupled and the system is nonlinear even when  $f$  is linear. For more information on (3.1) - (3.3), see [Scheidegger, 1974] and [Peaceman, 1977].

In order to be able to discuss the calculation, we present a flow chart for the highest level routines and the overall control flow, see Figure 3.1.

There are up to four distinct grids in this calculation. We identify each grid and explain its role. The most basic grid is the fixed hyperbolic grid. This is a two dimensional grid, and may be rectangular or curvilinear. It does not change with time, or only changes rarely. Thus if there is a fixed time independent flow which is known in advance to ap-

proximate the time dependent flow which is being sought as a solution to the equations

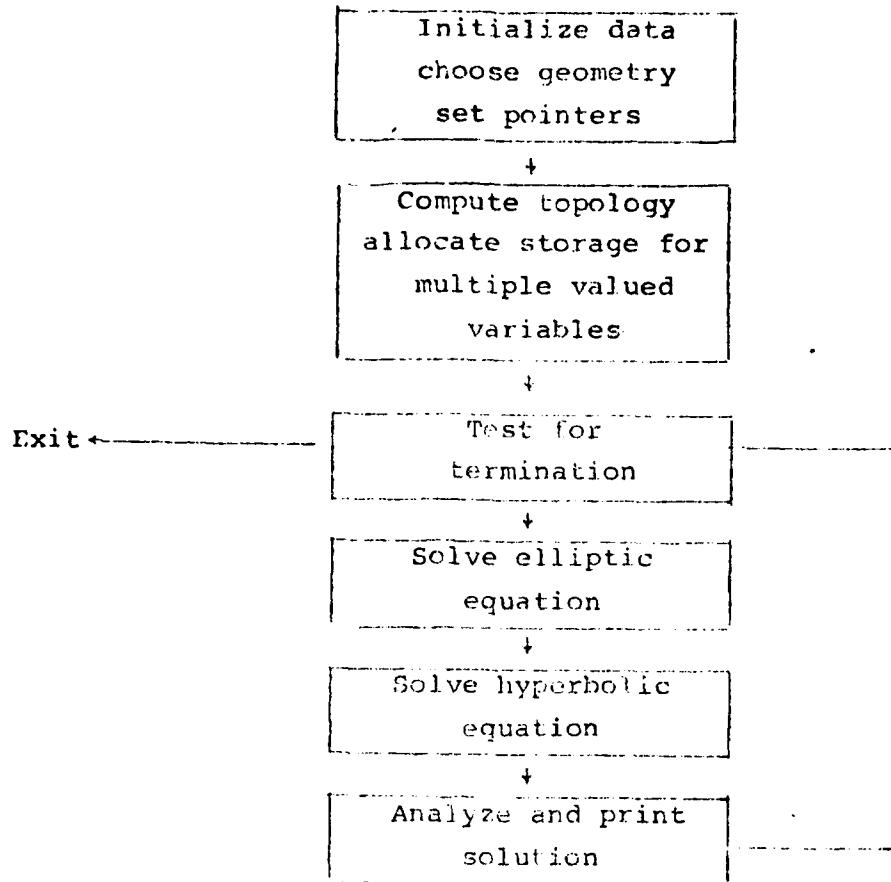


Figure 3.1 Flow Chart

(3.1)-(3.3), its stream and potential function can be used to construct a curvilinear grid. Otherwise a rectangular grid is used. The hyperbolic state variables are stored on this grid. Because the grid is fixed, the interpolation which results from remeshing is kept to a minimum, but not totally eliminated.

The track waves are described by a one dimensional time dependent and dynamically propagated grid. This grid is called the hyperbolic interface.

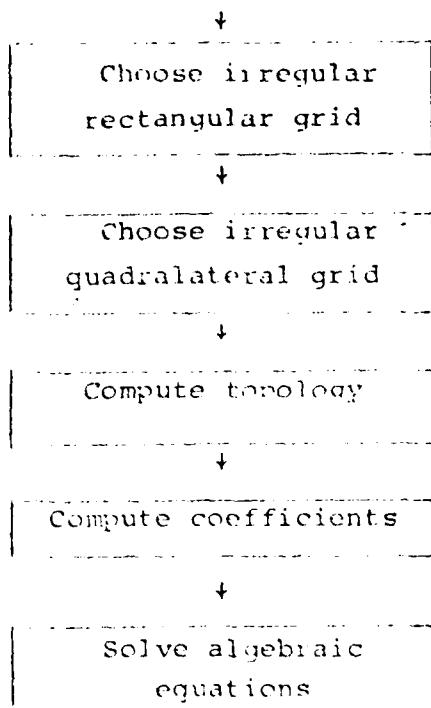


Figure 3.2 Flow Chart for solution of elliptic equation

Let  $\Lambda$  (e.g.  $[0,1] \times [0,1]$ ) be the region in which the equations are to be solved, and let  $F$  be the locus of the interface, thought of as a collection of curve segments. Then  $\Lambda \setminus F$  is not connected, but is a union of distinct connected components. For each hyperbolic mesh square, we also store topological information: which connected components of  $\Lambda \setminus F$  meet the mesh square. If the number of components  $n = n_{\text{comp}}$  is greater than one, then the basic hyperbolic state variables are multiple valued in this mesh square, and a distinct state is stored for each component meeting the mesh square.

In addition, the elliptic equation solver, especially the mesh alignment algorithm has its own two dimensional mesh where the pressure and velocity values are computed and in general it will have a slightly different one dimensional grid for representation of the interface.

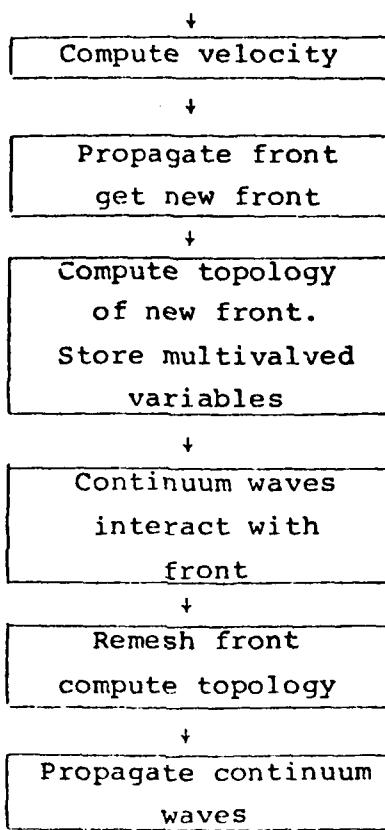


Figure 3.3. Flow Chart for solution of the hyperbolic equation.

A careful choice of the elliptic grid is the central virtue of the elliptic mesh alignment algorithm [McBryan, 1979]. First a nonuniform, but rectangular grid is chosen. The non-uniform spacing of grid lines is a simple one dimensional mesh refinement strategy. It permits a concentration of grid lines in regions of greatest interest, but because of its one dimensional character, is typically somewhat inefficient.

See Figure 3.4.

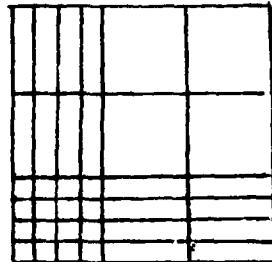


Figure 3.4. One dimensional mesh refinement.

Next this rectangular grid is distorted to a quadralateral grid near the front, by sliding nodes along grid lines until they intersect the front. Finally, the quadralateral grid is triangulated, being careful to pick the diagonal which follows the front if the front joins diagonally opposite nodes of a quadralateral.

Once discretized, the resulting algebraic system of equations is solved by a standard direct or iterative solver.

The method of mesh alignment appears to be superior to other methods of solving elliptic equations with strongly discontinuous coefficients across an irregular and distorted interface, especially where the derivate of the solution must be evaluated accurately along the discontinuity.

The propagation of the front uses an ordinary differential equation defined by the characteristic speed. The latter comes from a solution of the Riemann problem, using the states on the two sides of the front. (Recall that the hyperbolic states are multiple valued within a single mesh square, and have in particular distinct values ahead and behind the front.)

The computation of topology is necessary to determine the multiple valuedness of the updated hyperbolic states at the end of the time step. These calculations are kept at or near  $O(n)$  where  $n$  is the number of front points, and thus they do not contribute significantly to the overall computational time.

The interaction of continuum waves with the front is mediated by the same Riemann problem which determined the characteristic velocity of front. In fact this Riemann problem will in general contain  $n$  waves. The  $n-1$  waves other than the one being tracked are continuum waves which have been reflected off of the front or transmitted through it.

It is necessary to remesh the front from time to time, because interface points accumulate at some parts of the front and separate at other parts. The remeshing algorithms involve some degree of convex interpolation and thus both stabilize the front and degrade oscillations in it. Here we also check for changes of front topology (tangles). Eventually any new

topology (front crossings, tracked wave interactions and bifurcations) should be assimilated and propagated dynamically by the calculation, but these algorithms are not yet implemented.

The propagation of continuum waves uses values defined within a single connected component of  $\Delta F$ . Thus this is a - totally smooth, or untracked problem. State values from distinct sides of the tracked interface do not interact here. The operator split version [Chorin, 1976] of the uniform sampling method [Glimm, 1965] is used, but presumably a second order finite difference would also work and give superior results.

The organization is modular, and it is being revised to increase its modularity. Thus the problem dependent routines, such as the Riemann problem solver and isolated in a separate file, and can be easily changed, to change the code from one problem to another. Also portions of the code which have independent usefulness can be easily extracted and used out of context. Examples are the elliptic equation solver, the Riemann problem solver and the topology - interface package.

#### IV Applications

The ultimate scope of front tracking methods should be 3-d time dependent hydrodynamics calculations for problems with significant discontinuities and for which a priori knowledge of elementary wave interactions (e.g. the Riemann problem) is known. One proposed application is the Stephan problem. Although the temperature is continuous across the phase transition interface, the temperature gradients and tangential components of heat flux are in general discontinuous. Common experience (with melting ice, and with snowflakes) suggests that phase transition interfaces may be either stable or unstable. Another application is gas dynamics. The primary testing of the code has been in the context of petroleum reservoir simulation, and so we discuss this application in more detail.

One dimensional calculations.

In one dimension, the uniform sampling method [Glimm, 1965] gives excellent results because the correct structure of elementary waves and their interactions is built into the method. Tracking improves accuracy [Glimm, Marchesin and McBryan, 1980a]. Only for extremely stiff problems, such as flame propagation, is the extra accuracy likely to be worth the effort. The uniform sampling method has been tested successfully on a number of applications. Its use in petroleum problems began with [Concus and Proskurwoski, 1979]. Here it resolved fronts (water banks, i.e. shock waves) sharply without numerical diffusion or numerical instability. A 2x2 system modelling polymer injection (and a prototype for general surfactant recovery methods) was developed by [Isaacson, 1981] using the uniform sampling method. The mathematical interest in this model arises from the loss of strict hyperbolicity. In the region of coinciding wave speed, comparison calculations by finite difference methods were unable to resolve the true wave interactions, even on a very fine mesh. The engineering interest in controlling numerical diffusion lies in the case of surfactant recovery. The surfactant is expensive, and used only in thin layers. Its effect is nonlinear in the concentration, and is ineffective at low concentration. Thus too much or too little diffusion would give incorrect recovery results. The uniform sampling method was also applied to the flow in gas pipelines [Marchesin and Paes-Leme, 1981]. Here the e.g. by the opening or closing of valves. The uniform sampling method appears to be better than finite differences on this problem.

Two dimensions with operator splitting.

The extension of the uniform sampling method to two dimensions by operator splitting is not recommended in general. Negative results are due to [Collela, 1979 and Crandell and Majda, 1980]. In case the discontinuous waves are approximately parallel to the coordinate axis (and in some other special cases) satisfactory results can be obtained. In

[Glimm, Marchesin and McBryan, 1980b-1981] multiple fingers were resolved in a Taylor-Saffinan interface fingering instability, for a parameter range in which the instability was not too strong.

Two dimensions with tracking of discontinuities.

To overcome the serious limitations of x-y operator splitting, a front tracking code has been developed, as discussed in section III. It does not restrict the orientation or the topology of the front, and has been tested for mobility ratios up to 100, thus overcoming the principle restrictions of the operator splitting method. It also overcomes the main limitations of finite differences: numerical diffusion and grid orientation effects.

A statistical study of fingers.

Fingering of an interface is caused by a mismatch of the mobilities between two fluids. If the behind, or upstream fluid flows more easily, then an interface separating the two fluids and normal to the flow is unstable against formation of fingers. The instability is initiated by heterogeneity (which is certainly present in rock formations). The correct formulation and analysis of this problem is statistical. A preliminary study of the statistics [Glimm, Marchesin and McBryan, 1980b-1981] indicates that the rate of growth of fingers is independent of the heterogeneity and at least for parameter range which is typical of water flood problems, that areal heterogeneity does not affect recovery at breakthrough. (Channeling due to vertical variation of layers was not included in this study.) In general, the focus of a study of statistics of fingering should be to find relevant functionals of the solution which are either independent of the statistics in a simple and predictable fashion.

Deterministic fingers.

Choice of irregular Cauchy data is a deterministic method for computation of fingers. A number of calculations of this type were performed as part of the validation of the front tracking code, see [Glimm, Isaacson, Marchesin, McBryan, 1981]. Deterministic fingers were able to fit experimental data in tests up to mobility 5, but judging from preliminary results, agreement can probably be maintained for much larger mobility ratios.

Coarse grid calculations.

Most tests were performed 30x30 grid. Sample calculations on finer grids were also performed. For problems which are not too singular (and typical of a waterflood recovery process), reasonable results can be obtained from grids in the range 5x5 to 15x15. The ability to use coarse grids is essential for ultimate application to large scale problems.

Validation.

It is known that finite difference methods have severe mesh orientation problems on problems of the nature considered here. This means that a rotation of the grid by  $45^\circ$ , for example causes considerable difference in the computed solution. The reason, apparently, is that some orientations diminish the physical instability, while others may be neutral in effect or may enhance it. Tracking is intrinsically less grid dependent than the method of finite differences. Only small grid orientation effects were observed in the tracking calculations even for fairly singular parameters, and comparison was made to a grid which we believe to be neutral in its effect on the physical instability. Tests were also performed for convergence under mesh refinement and for agreement with experimental data. Further tests on finer grids are planned.

BIBLIOGRAPHY.

Chorin, A. (1976). Random choice solutions of hyperbolic systems. *J. Comp. Phys.* 22, 517-533.

Collela, P. (1979). An analysis of the effect of operator splitting and of the sampling procedure on the accuracy of Glimm's method. University of California Thesis, Berkeley.

Concus, P. and Proskurowski, W. (1979). Numerical solution of a nonlinear hyperbolic problem by a random choice method. *J. Comp. Phys.* 30, 153-166.

Courant, and Friedrichs, K. (1948). Supersonic flow and shock waves. Interscience, New York.

Crandell, M. and Madja, A. (1980). The method of fractional steps for conservation laws. *Num. Math.* 34, 285-314.

Dafermos, C. (1973). Solution of the Riemann problem for a class of hyperbolic systems of conservation laws by the viscosity method. *Arch. Rat. Mech. Anal.* 52, 1-9.

Dafermos, C. (1974). Structure of solutions of the Riemann problem for hyperbolic systems of conservation laws. *Arch. Rat. Mech. Anal.* 53, 203-217.

Dafermos, C. and DiPerna, R. (1973). The Riemann problem for certain classes of hyperbolic systems of conservation laws. *J. Diff. Eq.* 20, 90-114.

DiPerna, R. (1975). Singularities of solutions of nonlinear hyperbolic systems of conservation laws. *Arch. Rat. Mech. Anal.* 60, 75-100.

DiPerna, R. (1979). Uniqueness of solutions of hyperbolic conservation laws. *Indiana U. Math. J.* 28, 137-187.

Glimm, J. (1965). Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.* 18, 695-715.

Glimm, J., Isaacson, E., Marchesin, D., and McBryan, O. (1981). Front tracking for hyperbolic systems. *Adv. Appl. Math.* to appear.

Glimm, J. and Lax, P. (1970). Decay of solutions of systems of nonlinear conservation laws. *AMS Memoirs* 101.

Glimm, J., Marchesin, D. and McBryan, O. (1980a) Subgrid resolution of fluid discontinuities II. *J. Comp. Phys.* 37, 336-354.

Glimm, J., Marchesin, D. and McBryan, O. (1980b). Statistical fluid dynamics: Unstable fingers. *Comm. Math. Phys.* 74, 1-13.

Glimm, J., Marchesin, D. and McBryan, O. (1981). Unstable fingers for two phase flow. *Comm. Pure Appl. Math.* 34, 53-75.

Godunov, S. (1959). Difference methods for the numerical calculation of discontinuous solutions of the equations of fluid dynamics. *Math. Sb.* 47, 271-306 (in Russian).

Isaacson, E. (1981). Global solution of a Riemann problem for a nonstrictly hyperbolic system of conservation laws arising in enhanced oil recovery. Preprint.

Keyfitz, B. and Kranzer, H. (1980). A system of nonstrictly hyperbolic conservation laws arising in elasticity theory. *Arch. Rat. Mech. Anal.* 72, 219-241.

Keyfitz, B. and Kranzer, H. (1981). The Riemann problem for a class of hyperbolic conservation laws exhibiting a parabolic degeneracy. Preprint.

Lax, P. (1957). Hyperbolic systems of conservation laws II. *Comm. Pure Appl. Math.* 10, 537-566.

Liu, T.-P. (1974). The Riemann problem for general  $2 \times 2$  conservation laws. *Trans AMS* 199, 89-112.

Liu, T.-P. (1975). The Riemann problem for general system of conservation laws. *J. Diff. Eq.* 18, 218-234.

Liu, T.-P. (1977). The deterministic version of the Glimm scheme. *Comm. Math. Phys.* 57, 135-148.

Liu, T.-P. (1981). Admissible solutions of hyperbolic conservation laws. *Memoirs AMS*, 240.

Marchesin, D. and Paes-Leme, P. (1981). Shocks in gas pipelines. Rockefeller University preprint.

McBryan, O. (1980). Elliptic and hyperbolic interface refinement. In: *Boundary layers and interior layers -- computational and asymptotic methods*. J. Miller (Ed.) Boole Press, Dublin

Nishida, T. (1968). Global solution for an initial value problem of a quasilinear hyperbolic system. *Proc. Jap. Acad.* 44, 642-646.

Peaceman, D. (1977). Fundamentals of numerical reservoir simulation. Elsevier, New York.

Scheidegger, A. (1974). The physics of flow through porous media. University of Toronto Press, Toronto.

Temple, B. (1981). Global existence for a class of 2x2 nonlinear conservation laws with arbitrary Cauchy data. Rockefeller University preprint.

Wendroff, B. (1972). The Riemann problem for materials with a nonconvex equation of state. J. Math. Anal. Appl. 38, 454-466.

Supported in part by the NSF, grant PHY80-09179, by the ARO, grant DAAG29-79-C-1079 and the DOE, contract DEA-C02-76ER-03077.

The Rockefeller University  
Mathematics Department  
1230 York Avenue  
New York, N.Y. 10021

Courant Institute - NYU  
251 Mercer Street  
New York, N.Y. 10012

OVERTAKING OF SHOCK WAVES IN STEADY  
TWO-DIMENSIONAL SUPERSONIC FLOWS

by

Ling Hsiao \*  
Institute of Mathematics  
Academia Sinica of China  
Peking, China

and

(Visiting Professor, 1979-82)  
Division of Applied Mathematics  
Lefschetz Center for Dynamical Systems  
Brown University  
Providence, R. I. 02912

and

Tong Zhang  
Institute of Mathematics  
Academia Sinica of China  
Peking, China

June 1981

---

\* This research has been supported in part by the National Science Foundation under Grant #NSF-Eng. CME80-23824.

Talk presented by Ling Hsiao.

OVERTAKING OF SHOCK WAVES IN STEADY  
TWO-DIMENSIONAL SUPERSONIC FLOWS

by

Ling Hsiao and Tong Zhang

ABSTRACT

The purpose of the present paper is to study the overtaking of shock waves of the same family in a two-dimensional steady flow with polytropic gas.

It is proved that besides a transmitted shock and a contact discontinuity resulting from the overtaking of shocks of the same family, there is a reflected wave which is either a rarefaction wave or a shock. The criteria that determine whether the reflected wave is a shock or not is given in Theorems 1-3 in §3. The configuration of four shocks through one point is then presented when the reflected wave is a shock.

### 1. INTRODUCTION

There have been intense interests in the calculation of shock waves in the multi-dimensional gas flows. For the calculation of a single shock front, the Rankine-Hugoniot condition provides enough information to follow the shock front. However in an actual flow, more complicated wave patterns are involved. This is the case, for instance, when Mach stems appear in the flow around a body.

To calculate such a flow, it is helpful to understand analytically wave patterns involving interactions of shock fronts. It is particularly helpful when one uses the shock tracking technique to supplement an upwind difference scheme. Of course, the understanding of wave patterns is important in study of the qualitative behavior of the shock waves.

The purpose of the present paper is to study the overtaking of shock waves of the same family in a two-dimensional steady flow.

The steady plane flow (without viscosity) is described by the following system:

$$\left\{ \begin{array}{l} (\rho u)_x + (\rho v)_y = 0 \\ (\rho u^2 + p)_x + (\rho uv)_y = 0 \\ (\rho uv)_x + (\rho v^2 + p)_y = 0 \\ \left[ \rho u \left( h + \frac{u^2 + v^2}{2} \right) \right]_x + \left[ \rho v \left( h + \frac{u^2 + v^2}{2} \right) \right]_y = 0 \end{array} \right. \quad (1.1)$$

where  $\rho$  - density,  $p$  - pressure,  $(u, v)$  - velocity,  $h$  - enthalpy. This is in Eulerian coordinate for the flow without viscosity and external forces. The changes of state are adiabatic.

In this paper we consider polytropic gas, therefore

$$h = \frac{\gamma p}{(\gamma-1)\rho},$$

$\gamma > 1$  is adiabatic exponent. The flow is called supersonic if  $u^2 + v^2 > c^2$ ,  $c$  is sonic velocity,  $c^2 = \frac{\gamma p}{\rho}$  in present case. It is well-known that the system is hyperbolic when the flow is supersonic and there may be two kinds of shocks in the solution.

It is proved that besides a transmitted shock and a contact discontinuity resulting from the overtaking of shocks of the same family, there is a reflected wave which is either a rarefaction wave or a shock. The criteria that determine whether the reflected wave is a shock or not is given in Theorems 1-3 in §3. The configuration of four shocks through one point is then presented when the reflected wave is a shock.

Finally we discuss the interaction of a shock wave with a contact discontinuity (§4). Under the assumption that the magnitude of the shock is sufficiently weak the reflected wave is either a shock or a rarefaction wave, the criteria which determine what the reflected wave should be is given in theorem 4 in §4.

It should be pointed out that so far as the overtaking of shock waves is concerned there is an essential difference between steady

flow in two-dimensional and insteady flow in one-dimensional. For the latter, the reflected wave is always a simple centered wave no matter what the flow is, isentropic ([2]) or adiabatic (when  $\gamma \leq \frac{5}{3}$ , [1,3]). But for the former, the reflected wave may be a shock even if  $\gamma \leq \frac{5}{3}$ .

## 2. PRELIMINARY REMARKS

A self-similar solution of (1.1),  $(u, v, p, \rho) (x, y) = (u(\xi), v(\xi), p(\xi), \rho(\xi))$ ,  $\xi = \frac{y}{x}$ , satisfies the following system

$$\left( \begin{array}{cccc} v - \xi u & 0 & -\frac{\xi}{\rho} & 0 \\ 0 & v - \xi u & \frac{1}{\rho} & 0 \\ -\xi \rho & \rho & 0 & v - \xi u \\ 0 & 0 & v - \xi u & -c^2(v - \xi u) \end{array} \right) \begin{pmatrix} du \\ dv \\ dp \\ d\rho \end{pmatrix} = 0. \quad (2.1)$$

Let the determinant of the matrix of (2.1) be zero, it turns out

$$(v - \lambda u)^2 [(v - \lambda u)^2 - c^2(\lambda^2 + 1)] = 0, \quad \lambda = \xi, \quad (2.2)$$

which is called the characteristic equation of (1.1).

Corresponding to flow characteristic  $\lambda_0 = \frac{v}{u}$  which comes from the first factor of (2.2) there is two-dimensional manifold  $R_0$  in  $(u, v, p, \rho)$  space, namely

$$\begin{cases} p = \text{constant} \\ \frac{v}{u} = \text{constant} \end{cases} \quad (2.3)$$

Corresponding to wave characteristic

$$\lambda_i = \frac{uv \mp c \sqrt{u^2 + v^2 - c^2}}{u^2 - v^2}, \quad i = 1, 2 \quad (2.4)$$

which results from the second factor of (2.2) there is one-dimensional manifold  $R_i$ ,  $i = 1, 2$ , in  $(u, v, p, \rho)$  space which is defined by the following:

$$\begin{aligned} dp &= c^2 d\rho \\ du &= -\lambda_i dv \\ dp &= \rho(\lambda_i u - v) dv \end{aligned} \quad (2.5)$$

set  $w = \frac{v}{u}$ . It is easy to show that the projection of  $R_i$  into  $(p, w)$  plane is monotone

$$\frac{dw}{dp} = \mp \frac{\sqrt{c^2(u^2 + v^2 - c^2)}}{u^2 c^2 \rho} \quad (2.6)$$

here - (or +) corresponds to  $R_1$  (or  $R_2$ ) respectively.

A centered simple rarefaction wave  $\xi = \lambda_i (u, v, p, \rho)$  is determined by (2.5) and the requirement that the  $p$  value in wave front is greater than the  $p$  value in wave back.

It is well-known that any discontinuity in the solution of (1.1) has to be satisfied with the following Rakine-Hugoniot condition

$$\left\{ \begin{array}{l} \sigma[\rho u] = [\rho v] \\ \sigma[\rho u^2 + p] = [\rho uv] \\ \sigma[\rho uv] = [\rho_v^2 + p] \\ \sigma[\rho u(h + \frac{u^2 + v^2}{2})] = [\rho_v(h + \frac{u^2 + v^2}{2})] \end{array} \right. \quad (2.7)$$

where  $\sigma$  is the slope  $\frac{dy}{dx}$  of the discontinuity line, [ ] denotes the difference between the values on the right side and left side.

It is not difficult to prove that (2.7) is equivalent to the following:

$$\left| \begin{array}{cccc} \rho_0(v_0 - \sigma u_0) & 0 & -\sigma & 0 \\ 0 & \rho_0(v_0 - \sigma u_0) & 1 & 0 \\ -\sigma p & p & 0 & v_0 - \sigma u_0 \\ 0 & 0 & v_0 - \sigma u_0 & -\frac{c_0^2}{b}(v_0 - \sigma u_0) \end{array} \right| \left| \begin{array}{c} u - u_0 \\ v - v_0 \\ p - p_0 \\ \rho - \rho_0 \end{array} \right| = 0 \quad (2.8)$$

Here  $b = \frac{\gamma+1}{2} - \frac{\gamma-1}{2} \frac{p}{p_0}$ ,  $(u, v, p, \rho)$  and  $(u_0, v_0, p_0, \rho_0)$  are the states on the two sides of a discontinuity.

Let the determinant of the matrix of (2.8) be zero, it turns out that

$$(v_0 - \sigma u_0)^2 [(v_0 - \sigma u_0)^2 - \frac{p}{p_0} \frac{c_0^2}{b} (\sigma^2 + 1)] = 0 \quad (2.9)$$

Corresponding to  $\sigma_0 = \frac{v_0}{u_0}$  which comes from the first factor of (2.9), there is contact discontinuity  $R_0$ , namely,

$$\begin{cases} p = p_0 \\ \frac{v}{u} = \frac{v_0}{u_0} \end{cases} \quad (2.10)$$

corresponding to

$$\sigma_i = \frac{u_0 v_0 \mp \sqrt{\frac{c_0^2}{b} \frac{p}{p_0} (u_0^2 + v_0^2 - \frac{c_0^2}{b} \frac{p}{p_0})}}{u_0^2 - \frac{p}{p_0} \frac{c_0^2}{b}}, \quad i = 1, 2, \quad (2.11)$$

which results from the second factor of (2.9). There is one-dimensional manifold,  $i = 1, 2$ , in  $(u, v, p, \rho)$  space defined by

$$\begin{cases} p - p_0 = \frac{c_0^2}{b} (\rho - \rho_0) \\ u - u_0 = -\sigma_i (v - v_0) \\ \rho_0 (u_0 \sigma_i - v_0) (v - v_0) = (p - p_0). \end{cases} \quad (2.12)$$

For any given  $(u_0, v_0, p_0, \rho_0)$ , (2.12) determines a curve which passes through the given point and is denoted by  $S_i(0)$  corresponding to  $\sigma_i$ ,  $i = 1, 2$ .

A shock wave with slope  $\sigma_i$  is determined by (2.12) and the requirement that the  $p$  value on wave front is less than the  $p$  value on wave back.

Without loss of generality, we take  $v_0 = 0$ . It can be shown (see Appendix for details) that the projection of  $S_i(0)$  into  $(p, w)$  plane (still denote it by  $S_i(0)$ , Figure 2.1) has the following

expression for  $p \geq p_0$ :

$$\frac{dw}{dp} = \frac{\pm \sqrt{c^2 \cdot \frac{b' \rho}{b \rho_0} (bu_0^2 - c_0^2 \frac{\rho}{\rho_0})}}{2u^2 c^2 \rho} \left[ 1 + \frac{b \rho_0}{\rho} - \frac{\gamma+1}{2} \frac{c_0^2}{bu_0^2 - c_0^2 \frac{\rho}{\rho_0}} \cdot \frac{\rho - \rho_0}{\rho} \right]. \quad (2.13)$$

Here - (or +) corresponds to  $s_1(0)$  (or  $s_2(0)$ ) respectively,

$$b' = \frac{\gamma+1}{2} - \frac{\gamma-1}{2} \frac{\rho_0}{\rho}.$$

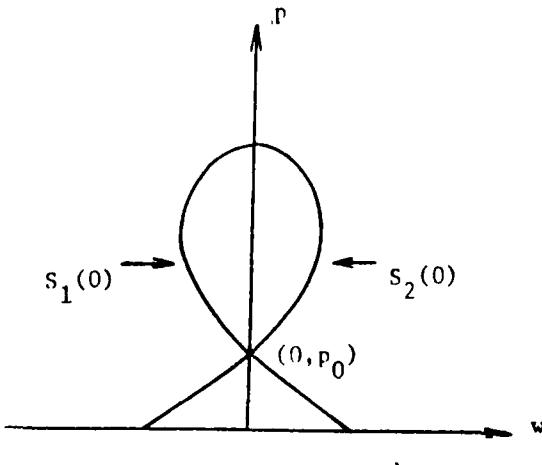


Figure 2.1

Set  $\frac{\rho}{\rho_0} = t$ , (2.13) can be rewritten as

$$\frac{dw}{dp} = \frac{\pm \sqrt{c^2 \frac{b't}{b} (bu_0^2 - c_0^2 t)}}{2u^2 c^2 p_0 t} \left[ 1 + \frac{b}{t} - \frac{\gamma+1}{2} \frac{c_0^2}{bu_0^2 - c_0^2 t} \cdot \frac{t-1}{t} \right]. \quad (2.14)$$

Here  $b(t) = \frac{\gamma+1}{2} - \frac{\gamma-1}{2} t$ ,  $b'(t) = \frac{\gamma+1}{2} - \frac{\gamma-1}{2t}$ .

Denote by  $t = t'$  the zero of  $bu_0^2 - c_0^2 t$ , i.e.,

$$t' = \frac{\frac{\gamma+1}{2} u_0^2}{\frac{\gamma-1}{2} u_0^2 + c_0^2}.$$

It is obvious that  $1 < t' < \frac{\gamma+1}{\gamma-1}$  for  $u_0^2 > c_0^2$ . Therefore (2.14) makes sense for  $1 \leq t \leq t'$  and represents the part of the curve  $S_1(0)$  which corresponds to  $p \geq p_0$ .

It is easy to see that there exists unique point  $t = \bar{t}$  in the interval  $(1, t')$  which vanishes (2.14). Consequently,  $S_1(0)$  is monotone in  $[1, \bar{t}]$ .

On account of (2.11), (2.12), it can be shown that the following expression holds along  $S_2(0)$ :

$$u^2 + v^2 - c^2 = c_0^2 \{M_0^2 - T(t)\} \quad (2.15)$$

where  $M_0$  is the Mach number of the state  $(u_0, v_0, p_0, \rho_0)$  and

$$T(t) = \frac{2t^2 + (\gamma+1)(t-1)}{t[(\gamma+1) - (\gamma-1)t]}.$$

It is not difficult to prove that there exists a unique value  $t = t^*$  in  $(1, \frac{\gamma+1}{\gamma-1})$  with  $t^* < \bar{t}$ ,

$$M_0^2 - T(t) = 0 \quad \text{at } t = t^*$$

$$\text{and } M_0^2 - T(t) > 0 \quad \text{for } 1 \leq t < t^*.$$

Therefore, only those points of  $S_2(0)$  which correspond to  $1 \leq t < t^*$  will be used in the following sections since we are concerned with supersonic flows.

### 3. THE OVERTAKING OF SHOCK WAVES

We only consider the overtaking of two shocks  $S_2^{(1)}$  and  $S_2^{(2)}$  of the second kind in this section. (The overtaking of two shocks of the first kind is treated in an analogous way). Denote the wave front state and wave back state of the first shock  $S_2^{(1)}$  by  $(u_0 v_0 p_0 \rho_0)$ , the (0) state, and the (1) state, respectively. Without loss of generality, we take  $v_0 = 0$ ,  $u_0 > 0$  (Figure 3.1). Denote the wave back state of the overtaken shock  $S_2^{(2)}$  by  $(u_2 v_2 p_2 \rho_2)$ , the (2) state, and Q the point of intersection of  $S_2^{(1)}$  and  $S_2^{(2)}$ .

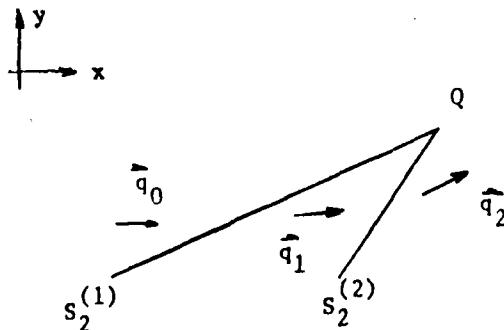


Figure 3.1

Obviously,

$$u_0^2 > c_0^2, \quad q_1^2 > c_1^2, \quad (q_1^2 = u_1^2 + v_1^2)$$

$$\frac{v_2}{u_2} > \frac{v_1}{u_1} > 0, \quad p_2 > p_1 > p_0.$$

The aim is to construct a solution which consists of a centered simple rarefaction wave and shock waves centered at  $Q$  and separated by constant states. We also want to give criteria to determine the configuration of the solution.

Consider the curves in  $(p, w)$  plane. We know that (1) state is on the shock polar curve  $S_2(0)$  and (2) state is on the shock polar curve  $S_2(1)$ . If (2) state is inside of  $S_2(0)$ , (Figure 3.2), then there exist two states  $(u_3, v_3, p_3)$  and  $(\bar{u}_3, \bar{v}_3, \bar{p}_3)$  satisfying

$$\frac{v_3}{u_3} = \frac{\bar{v}_3}{\bar{u}_3} \quad \text{and} \quad p_3 = \bar{p}_3$$

such that (3) state is on the curve  $R_1(2)$  and (3) state is on the curve  $S_2(0)$  respectively.

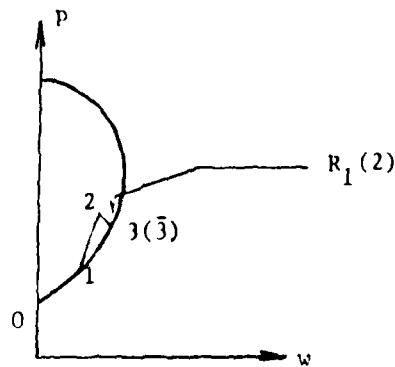


Figure 3.2

In this case, (see § 2), we can construct the solution which consists of a shock  $S_2$  with wave front state  $(u_0 v_0 p_0 \rho_0)$  and wave back state  $(\bar{u}_3 \bar{v}_3 \bar{p}_3 \bar{\rho}_3)$ , a centered simple rarefaction wave  $R_1$  with wave front state  $(u_2 v_2 p_2 \rho_2)$  and wave back state  $(u_3 v_3 p_3 \rho_3)$ , and a contact discontinuity  $T$  (Figure 3.3).

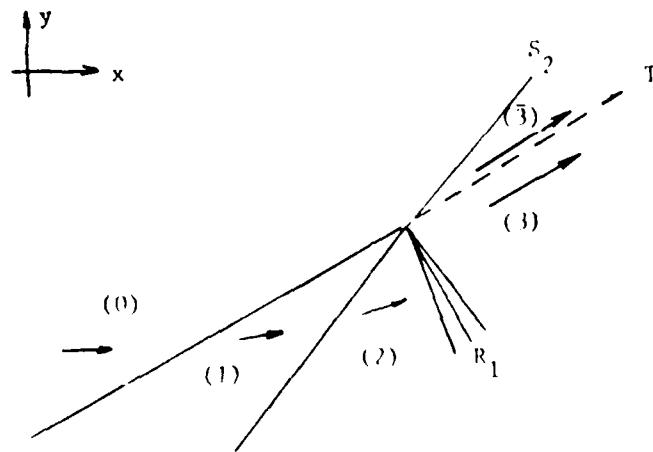
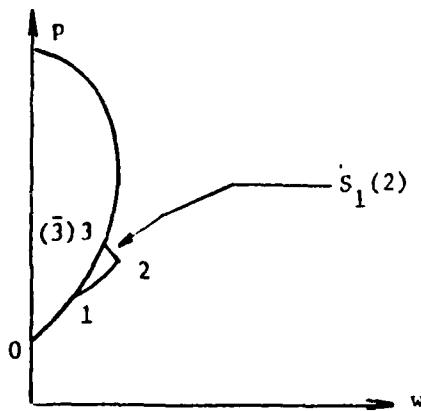
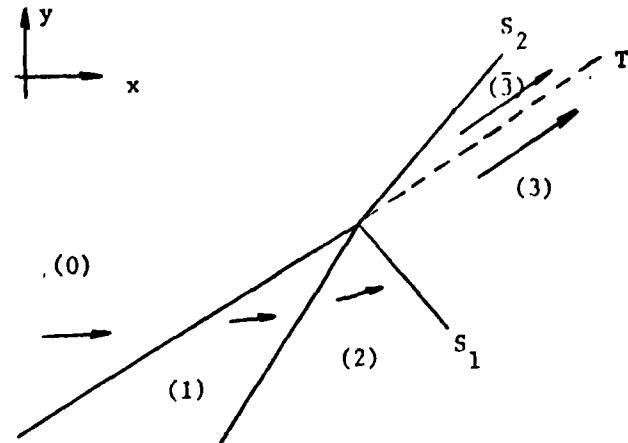


Figure 3.3

If (2) state is outside of the curve  $S_2(0)$  (Figure 3.4) then

there exist two states  $\underline{3}$  and  $\bar{3}$  satisfying  $\frac{v_3}{u_3} = \frac{\bar{v}_3}{\bar{u}_3}$  and  $p_3 = \bar{p}_3$

such that (3) state is on the curve  $S_1(2)$  and ( $\bar{3}$ ) state is on the curve  $S_2(0)$ . (The states  $\underline{3}$  and  $\bar{3}$  always exist when  $|p_2 - p_1|$  is small). When this happens the solution consists of a shock  $S_2$  with  $(u_0 v_0 p_0 \rho_0)$  and  $(\bar{u}_3 \bar{v}_3 \bar{p}_3 \bar{\rho}_3)$  as the wave front and wave back state, shock  $S_1$  with  $(u_2 v_2 p_2 \rho_2)$  and  $(u_3 v_3 p_3 \rho_3)$  as the wave front and wave back state, contact discontinuity  $T$  (Figure 3.5).

Figure 3.4Figure 3.5

So, whether the reflected wave is a shock or centered simple wave is reduced to whether (2) state is outside or inside of the curve  $S_2(0)$ .

It is well-known that the curve  $S_2(1)$  and  $R_2(1)$  are tangent at point (1) up to the second order. Therefore, it suffices to consider the relative position between  $R_2(1)$  and  $S_2(0)$  instead of the relative position between  $S_2(1)$  and  $S_2(0)$  when  $|p_2 - p_1|$  is sufficiently small.

Let  $H$  be the slope of the projection of  $S_2(0)$  into  $(p, w)$  plane, and  $A$  the slope of the projection of  $R_2(1)$  into  $(p, w)$  plane. We consider the sign of  $H-A$  along the curve  $S_2(0)$  when (1) state satisfies the following condition:

$$\frac{\rho_1}{\rho_0} < t^* \quad (3.1)$$

Since  $H > 0$  and  $A > 0$  when  $1 \leq t \leq t^*$ , we will study the

sign of  $H^2 - A^2$  instead of  $H-A$ , along the curve  $s_2(0)$ .

Direct calculations using (2.6) and (2.13) yields

$$H^2 - A^2 = \frac{1}{u^4 c^2 p^2} \left\{ \frac{b't}{b} (bu_0^2 - c_0^2 t) \left[ \frac{2(t+b)(bu_0^2 - c_0^2 t) - (\gamma+1)c_0^2(t-1)}{4t(bu_0^2 - c_0^2 t)} \right]^2 - \frac{u_0^2 bt - c_0^2(t^2 - 1 + b't)}{bt} \right\} \quad (3.2)$$

and so the sign of  $H^2 - A^2$  is the same as that of the function

$G(M_0^2, t)$ , where

$$G(M_0^2, t) = b^2 [(3-\gamma)^2 t - (\gamma^2 - 1)] M_0^4 - 2b [(3-\gamma)^2 t^2 - 2(\gamma+1)(\gamma-2)t + (\gamma^2 - 1)] M_0^2 + [(3-\gamma)^2 t^3 + (\gamma+1)(7-3\gamma)t^2 + (\gamma+1)(3\gamma-1)t - (\gamma^2 - 1)]$$

$$M_0^2 = \frac{u_0^2}{c_0^2} .$$

Now we discuss the sign of  $G(M_0^2, t)$  in the region  $\Pi$ :  $1 \leq t \leq t^*$ ,

$$M_0^2 \geq 1 \}.$$

$$\underline{\text{Case I: }} 1 < \gamma \leq \frac{5}{3}$$

Obviously,  $(3-\gamma)^2 t - (\gamma^2 - 1) > 0$  for  $t > 1$  in this case. On account

of  $G > 0$  when  $M_0^2 = T(t)$ ;  $G > 0$  when  $M_0^2 \rightarrow \infty$ ;  $\frac{\partial^2 G}{\partial(M_0^2)^2} > 0$  and

$$\min_{M_0^2 \geq T(t)} G(M_0^2, t) = \frac{-16(\gamma^2 - 1)t}{(3-\gamma)^2 t - (\gamma^2 - 1)} < 0 \quad \text{for } t \in [1, \frac{\gamma+1}{\gamma-1}) \quad \text{it can be shown}$$

that for any given  $t \in [1, \frac{\gamma+1}{\gamma-1})$  there exist  $M_{0_i}^2(t)$ ,  $i = 1, 2$ , such

that  $T(t) < M_{0_1}^2(t) < M_{0_2}^2(t) < \infty$ ,  $G(M_{0_i}^2(t), t) \geq 0$  and

$G \geq 0$  when  $T(t) < M_0^2 < M_{0_1}^2(t)$ ;

$G < 0$  when  $M_{0_1}^2(t) < M_0^2 < M_{0_2}^2(t)$ ;

$G > 0$  when  $M_0^2 > M_{0_2}^2(t)$ .

Furthermore,

$$M_{0_i}^2(t) = \frac{(3-\gamma)^2 t^2 - 2(\gamma+1)(\gamma-2)t + (\gamma^2-1) \mp 4\sqrt{(\gamma^2-1)t}}{b[(3-\gamma)^2 t - (\gamma^2-1)]} \quad (3.3)$$

here  $i = 1$  (or 2) corresponds to that one with minus (or plus) in front of the term  $\sqrt{(\gamma^2-1)t}$ .

Proposition 3.1.  $M_0^2 = M_{0_1}^2(t)$  is a monotone function of  $t$  on  $[1, \frac{\gamma+1}{\gamma-1})$ ;  $M_0^2 = M_{0_2}^2(t)$  is a convex function of  $t$  which attains minimum  $M_{0_m}^2$  at  $t = \tilde{t}$  on  $[1, \frac{\gamma+1}{\gamma-1})$ .

Proof. From (3.3),  $i = 1$ , one gets

$$\frac{dM_{0_1}^2(t)}{dt} = \frac{M(t) - \sqrt{\frac{\gamma^2-1}{t}} N(t)}{b^2 [(3-\gamma)^2 t - (\gamma^2-1)]^2} \quad (3.4)$$

where

$$M(t) = (\gamma+1)[(3-\gamma)^3 t^2 - 2(3-\gamma)^2 (\gamma-1)t + (\gamma^2-1)(3\gamma-7)] \quad (3.5)$$

$$N(t) = 3(\gamma-1)(3-\gamma)^2 t^2 - 2(\gamma+1)(\gamma^2-4\gamma+5)t - (\gamma+1)^2(\gamma-1) \quad (3.6)$$

Let  $M(t) = \sqrt{\frac{Y^2-1}{t}} N(t) = \phi(t)$ , we will show  $\phi(t) > 0$ .

Noticing that  $\phi(1) = (Y+1)\alpha(Y) + \sqrt{Y^2-1}\beta(Y)$ ,  $\beta(Y) > 0$  and

$$\alpha(Y) = 16(Y-Y^*)(Y-\tilde{Y}), \text{ where}$$

$$\alpha(Y) = 4(4Y^2-15Y+13)$$

$$\beta(Y) = 4(4Y^2-11Y+9)$$

$$\frac{5}{4} < Y^* < \frac{5}{3}, \quad 2 < \tilde{Y} < 3,$$

it is easy to see that  $\phi(1) > 0$  for  $1 < Y < \frac{5}{4}$  and  $\phi(1) > 0$  if and only if  $\Gamma(Y) > 0$  for  $\frac{5}{4} < Y \leq \frac{5}{3}$ , where

$$\Gamma(Y) = 72(Y - \frac{5}{3})^2(Y - \frac{5}{4}).$$

Thus  $\phi(1) > 0$  for  $1 < Y \leq \frac{5}{3}$ .

Similarly we know that  $\phi'(t) > 0$  ( $t \geq 1$ )  $\Leftrightarrow \tilde{\phi}(x) > 0$  ( $x \geq 1$ )

(let  $\sqrt{t} = x$ ) where

$$\begin{aligned} \tilde{\phi}(x) = 4(Y+1)(3-Y)^3x^5 - 9\sqrt{Y^2-1}(Y-1)(3-Y)^2x^4 - 4(Y^2-1)(3-Y)^2x^3 + \\ 2\sqrt{Y^2-1}(Y+1)(Y^2-4Y+5)x^2 - \sqrt{Y^2-1}(Y+1)^2(Y-1) \end{aligned}$$

and the following inequalities

$$(Y+1)(3-Y) \geq 4\sqrt{Y^2-1}(Y+1)$$

$$3-Y \geq 2(Y-1)$$

$$8(Y^2-4Y+5) > 5(Y^2-1)$$

$$2(Y+1)(Y^2-4Y+1) > 5(Y-1)(3-Y)^2,$$

and so  $\tilde{\phi}(x) > 0$  for  $x \geq 1$ , we have  $\phi'(t) > 0$  for  $t \geq 1$  which

together with  $\phi(1) > 0$  imply  $\phi(t) > 0$  for  $t \geq 1$ . Consequently,

$M_0^2 = M_{01}^2(t)$  is a monotone function of  $t$  on  $[1, \frac{\gamma+1}{\gamma-1}]$ .

From (3.3),  $i = 2$ , one gets

$$\frac{dM_{02}^2(t)}{dt} = \frac{M(t) + \sqrt{\frac{\gamma^2-1}{t}} N(t)}{b^2[(3-\gamma)^2 t - (\gamma^2-1)]^2}. \quad (3.7)$$

$$\text{Let } M(t) + \sqrt{\frac{\gamma^2-1}{t}} N(t) = \Psi(t). \quad (3.8)$$

It can be shown, in the similar way as the above, that

$$\underline{\Psi}(1) < 0 \quad (3.9)$$

On account of  $\Psi(\frac{\gamma+1}{\gamma-1}) = 16(\frac{\gamma+1}{\gamma-1})^2(2-\gamma) > 0$ , (3.9) and

$$\underline{\Psi}''(t) = 2(\gamma+1)(3-\gamma)^3 + \frac{\sqrt{\gamma^2-1}}{4} t^{-\frac{5}{2}} \tilde{\Psi}(t) > 0 \quad (3.10)$$

$$\text{where } \tilde{\Psi}(t) = 9(\gamma-1)(3-\gamma)^2 t^2 + 2(\gamma+1)(\gamma^2 - 4\gamma + 5)t - 3(\gamma+1)^2(\gamma-1), \quad (3.11)$$

it turns out that  $M_0^2 = M_{02}^2(t)$  is convex function of  $t$  and there exists

a value  $t = \tilde{t}$  on  $[1, \frac{\gamma+1}{\gamma-1}]$  such that  $M_0^2 = M_{02}^2(t)$  takes minimum  $M_{0m}^2$  at  $t = \tilde{t}$ . This completes the proof of proposition 3.1.

By using the proposition 3.1, we get the distribution of the sign of  $G(M_0^2, t)$  in the region  $\Pi$  as presented in Figure 3.6. Here

$$M_{01}^2 = \frac{2(3-\gamma) - 2\sqrt{\gamma^2-1}}{5-3\gamma} \quad (3.12)$$

and

$$M_{02}^2 = \frac{2(3-\gamma) + 2\sqrt{\gamma^2-1}}{5-3\gamma}. \quad (3.13)$$

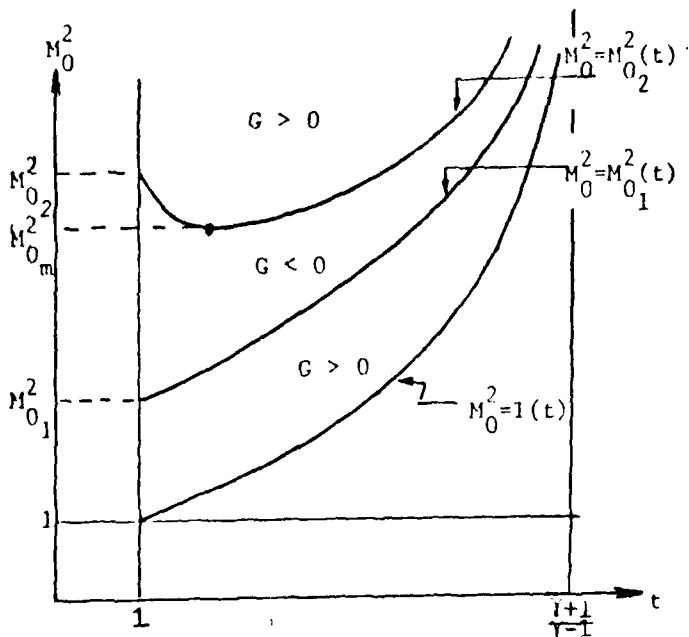


Figure 3.6

Case II:  $\frac{5}{3} < Y < 2$

It is obvious that there exists  $\bar{t} \in (1, \frac{Y+1}{Y-1})$  such that  $(3-Y)^2 t - (Y^2 - 1) < 0$  for  $1 \leq t \leq \bar{t}$  and  $(3-Y)^2 t - (Y^2 - 1) > 0$  for  $t > \bar{t}$ .

Since  $G > 0$  when  $M_0^2 = T(t)$ ;  $G < 0$  when  $M_0^2 \rightarrow \infty$  and  $\frac{\partial^2 G}{\partial (M_0^2)^2} < 0$ , it can be shown that for any  $t$ ,  $1 \leq t < \bar{t}$ , there exists a unique  $M_0^2(t)$  such that  $T(t) < M_0^2(t) < \infty$ ,  $G(M_0^2(t), t) \equiv 0$  and

$G > 0$  when  $T(t) < M_0^2 < M_0^2(t)$ ;

$G < 0$  when  $M_0^2 > M_0^2(t)$ ,

where  $M_{0_1}^2(t)$  has the same expression as (3.3) for  $i = 1$ .

For any given  $t$  with  $\bar{t} < t < \frac{\gamma+1}{\gamma-1}$ , it is similar to the case I that there exist  $M_{0_i}^2(t)$ ,  $i = 1, 2$  (to have the same expression (3.3)) such that

$$G > 0 \text{ when } T(t) < M_{0_1}^2 < M_{0_2}^2(t);$$

$$G < 0 \text{ when } M_{0_1}^2(t) < M_{0_2}^2 < M_{0_2}^2(t);$$

$$G > 0 \text{ when } M_{0_2}^2 > M_{0_1}^2(t).$$

Furthermore, we have the following proposition.

Proposition 3.2.  $M_0^2 = M_{0_1}^2(t)$  is a monotone function of  $t$  on

$[1, \frac{\gamma+1}{\gamma-1}]$ ;  $M_0^2 = M_{0_2}^2(t)$  is convex function defined on  $(\bar{t}, \frac{\gamma+1}{\gamma-1})$  with

$t = \bar{t}$  as its asymptote and takes minimum  $M_{0_m}^2$  at  $t = \tilde{t}$ .

Proof. Let  $M(t) = \sqrt{\frac{\gamma^2-1}{t}} N(t) = \Phi(t)$ , it is easy to check that

$M_0^2 = M_{0_1}^2(t)$  is smooth at  $t = \bar{t}$  with positive derivative. Thus,

in order to prove the first part of the proposition it suffices to prove

that  $\Phi(t) \geq 0$  on  $[1, \frac{\gamma+1}{\gamma-1}]$  which is a consequence of the following

identities:

$$\Phi(t) = 0 \text{ and } \Phi'(t) = 0 \text{ at } t = \bar{t}$$

and

$$\Phi''(t) = \Phi_1(t) + \frac{\sqrt{\gamma^2-1}}{4} t^{-\frac{5}{2}} \Phi_2(t) > 0$$

where

$$\Phi_1(t) = 2(\gamma+1)(3-\gamma)^3 - \frac{13}{4} \cdot \sqrt{\frac{\gamma^2-1}{t}} (\gamma-1)(3-\gamma)^2$$

and

$$\Phi_2(t) = 4(\gamma-1)(3-\gamma)^2 t^2 - 2(\gamma+1)(\gamma^2-4\gamma+5)t + 3(\gamma+1)^2(\gamma-1).$$

Clearly,  $t = \bar{t}$  is the asymptote of  $M_0^2 = M_{0_2}^2(t)$ .

From (3.7), (3.8), (3.9) and (3.10), it is easily seen that

$$\underline{\psi}(t) \rightarrow \frac{16(\gamma-2)(\gamma^2-1)(\gamma+1)}{3-\gamma} < 0 \quad \text{when } t \rightarrow \bar{t} + 0$$

$$\underline{\psi}(t) > 0 \quad \text{when } t \rightarrow \frac{\gamma+1}{\gamma-1}$$

and  $\underline{\psi}''(t) > 0$ .

Therefore there exists a unique  $t = \tilde{t} \in (\bar{t}, \frac{\gamma+1}{\gamma-1})$  with  $\underline{\psi}(t) = 0$ . In other words,  $M_0^2 = M_{0_2}^2(t)$  is convex with minimum  $M_{0_m}^2$  at  $t = \tilde{t}$ . This completes the proof of proposition 3.2.

By using proposition 3.2, we get the distribution of the sign of  $G(M_0^2, t)$  in the region  $\Pi$  as in Figure 3.7.

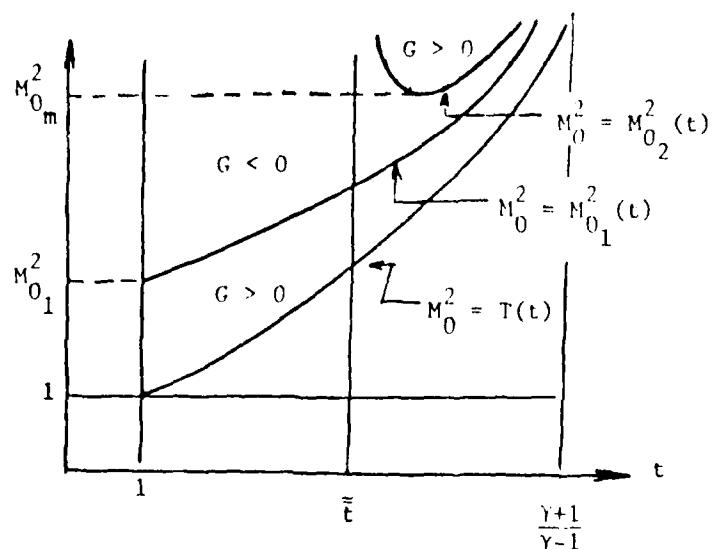


Figure 3.7

Case III.  $\gamma \geq 2$ 

Obviously,  $(3-\gamma)^2 t - (\gamma^2-1) < 0$  for  $1 \leq t < \frac{\gamma+1}{\gamma-1}$  in this case.

In the similar way as  $t < \bar{t}$  case in II, it can be shown that for any given  $t \in [1, \frac{\gamma+1}{\gamma-1})$  there exists unique  $M_{0_1}^2(t)$  (the same expression of (3.3) for  $i = 1$ ) such that  $G(M_{0_1}^2(t), t) \equiv 0$  and

$$G > 0 \quad \text{when } T(t) < M_{0_1}^2 < M_{0_1}^2(t);$$

$$G < 0 \quad \text{when } M_{0_1}^2 > M_{0_1}^2(t).$$

Furthermore, we have

Proposition 3.3.  $M_{0_1}^2(t)$  is a monotone function of  $t$  on  $[1, \frac{\gamma+1}{\gamma-1})$ .

Proof. On account of  $N(t) < 0$  at  $t = 1$ ,  $N(t) < 0$  at  $t = \frac{\gamma+1}{\gamma-1}$  and

$N''(t) > 0$ , it follows that  $N(t) < 0$  on  $[1, \frac{\gamma+1}{\gamma-1})$ , therefore

$$M(t) - \sqrt{\frac{\gamma^2-1}{t}} N(t) > M(t) - (\gamma-1)N(t) = 4(\gamma-2)B(t) \quad \text{on } [1, \frac{\gamma+1}{\gamma-1}),$$

here  $B(t) = (\gamma+1)^2(\gamma-1) + (\gamma^2-1)t - \gamma(3-\gamma)^2t^2$ . It is easily seen that  $B(t) > 0$  and the proposition is proved.

By the proposition (3.3) we obtain the distribution of the sign of  $G(M_{0_1}^2, t)$  in the region  $\Pi$  as presented in Figure 3.8.

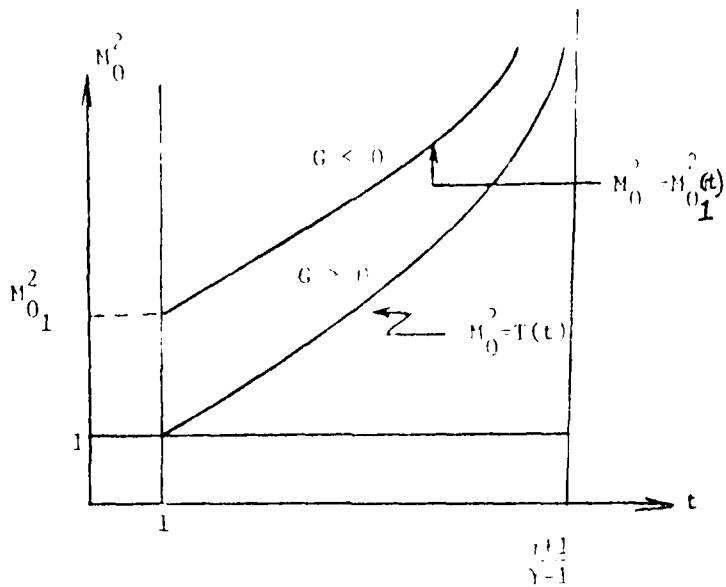


Figure 3.8

In view of these propositions, it ends up that in addition to a transmitted shock and a contact discontinuity, the result of overtaking of shock waves involves a reflected wave which is either a shock or a rarefaction wave. The criteria to judge the configuration are given in following theorems.

Theorem 3.1 When  $1 < Y \leq \frac{5}{3}$  and the overtaken shock  $s_2^{(2)}$  is sufficiently weak, there exist constants  $M_{0_i}^2$ ,  $i = 1, 2$ , which are given in (3.12), (3.13) and  $M_{0_m}^2$  which is determined by  $M_{0_m}^2 = M_{0_2}^2(\tilde{t})$  (see

Proposition 3.1) such that

If  $1 < M_0^2 < M_{0_1}^2$  ( $M_0^2$  is the Mach number of the (0) state), the

reflected wave is the first kind of centered simple wave (Figure 3.9).

If  $M_{0_1}^2 < M_0^2 < M_{0_m}^2$ , whether the reflected wave is a shock or not depends on the magnitude  $\frac{\rho_1}{\rho_0}$  of the first shock  $s_2^{(1)}$ . There exists

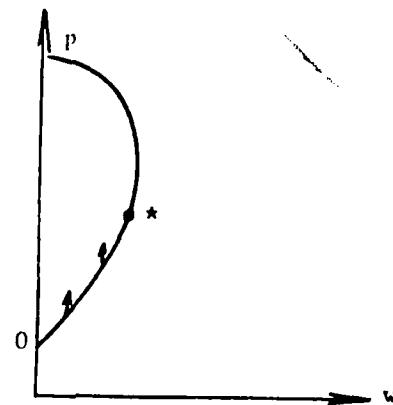


Figure 3.9

unique  $t_{M_0}^1$  such that the reflected wave is the first kind of shock

when  $1 < \frac{\rho_1}{\rho_0} < t_{M_0}^1$ ; the reflected wave is the first kind of centered

simple wave when  $t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^*$  (see Figure 3.10). The value  $t_{M_0}^1$  is determined by  $M_0^2 = M_{0_1}^2 (t_{M_0}^1)$ .

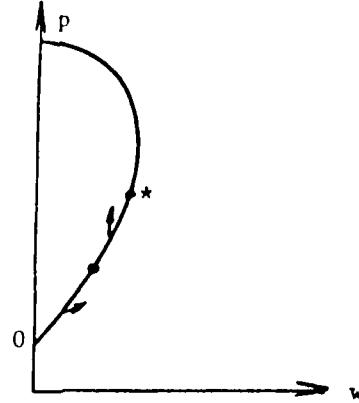


Figure 3.10

If  $M_{0_m}^2 < M_0^2 < M_{0_2}^2$ , there exist three values  $t_{M_0}^2 < t_{M_0}^{2'} < t_{M_0}^1$ ,

$t_{M_0}^2$  and  $t_{M_0}^{2'}$  are determined by  $M_0^2 = M_{0_2}^2 (t)$  and  $t_{M_0}^1$  by  $M_0^2 = M_{0_1}^2 (t)$ ,

such that the reflected wave is the first kind of shock when  $1 \leq \frac{\rho_1}{\rho_0} < t_{M_0}^2$ ;

the reflected wave is the first kind of centered simple wave when

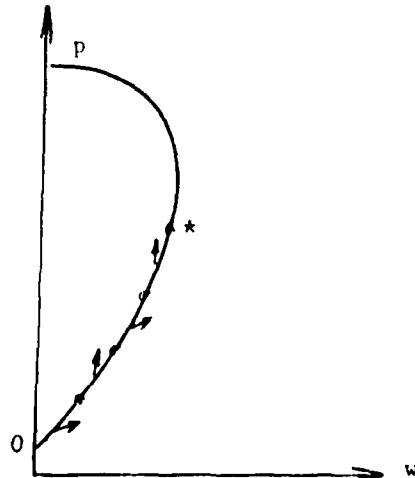
$$t_{M_0}^2 < \frac{\rho_1}{\rho_0} < t_{M_0}^{2'} ;$$

the reflected wave is the first kind of shock when  $t_{M_0}^{2'} < \frac{\rho_1}{\rho_0} < t_{M_0}^1$  ;

the reflected wave is the first kind of centered simple wave when

$$t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^* \text{ (see Figure 3.11).}$$

Figure 3.11



If  $M_0^2 > M_{02}^2$ , there exist two values  $t_{M_0}^i$ ,  $i = 1, 2$ , which are determined by  $M_0^2 = M_{0i}^2(t)$  respectively, such that

the reflected wave is the first kind of centered simple wave when

$$1 \leq \frac{\rho_1}{\rho_0} < t_{M_0}^2 ;$$

the reflected wave is the first kind of shock when  $t_{M_0}^2 < \frac{\rho_1}{\rho_0} < t_{M_0}^1$  ;

the reflected wave is the first kind of centered simple wave when

$$t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^* \text{ (see Figure 3.12).}$$

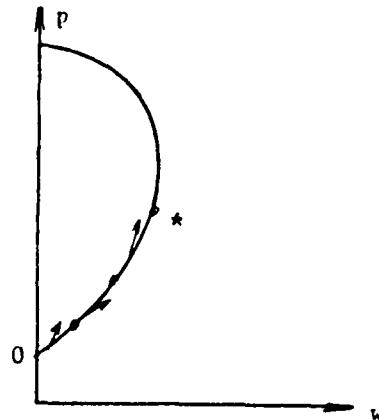
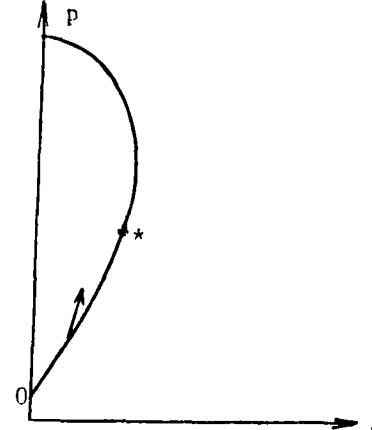


Figure 3.12

Theorem 3.2. When  $\frac{5}{3} < \gamma < 2$  and the overtaken shock  $s_2^{(2)}$  is sufficiently weak there exist constants  $M_{0_1}^2$  and  $M_{0_m}^2$  (defined as in Theorem 3.1) with the following properties:

If  $M_0^2 < M_{0_1}^2$ , the reflected wave is the first kind of centered simple wave (Figure 3.13).

Figure 3.13



If  $M_{0_1}^2 < M_0^2 < M_{0_m}^2$ , there exists a unique  $t_{M_0}^1$  such that the reflected wave is the first kind of shock when  $1 \leq \frac{\rho_1}{\rho_0} < t_{M_0}^1$ ;

the reflected wave is the first kind of centered simple wave when

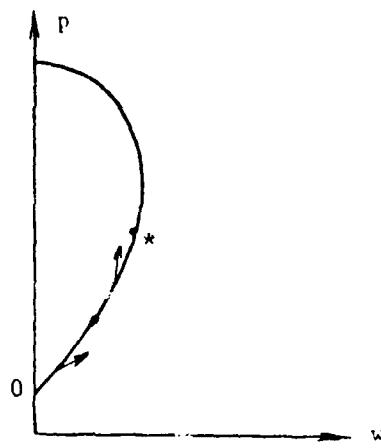


Figure 3.14

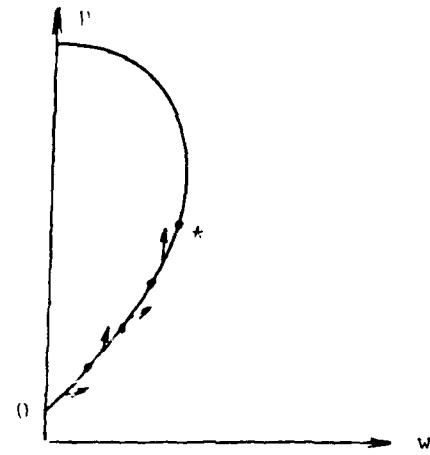


Figure 3.15

$$t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^* \quad (\text{Figure 3.14}).$$

If  $M_0^2 > M_m^2$ , there exist three values  $t_{M_0}^2 < t_{M_0}^{2'} < t_{M_0}^1$  determined

in the same way as Theorem 3.1 such that (Figure 3.15)

the reflected wave is the first kind of shock when  $1 \leq \frac{\rho_1}{\rho_0} < t_{M_0}^2$ ;

the reflected wave is the first kind of centered simple wave when

$$t_{M_0}^2 < \frac{\rho_1}{\rho_0} < t_{M_0}^{2'};$$

the reflected wave is the first kind of shock when

$$t_{M_0}^{2'} < \frac{\rho_1}{\rho_0} < t_{M_0}^1;$$

the reflected wave is the first kind of centered simple wave when

$$t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^*.$$

Theorem 3.3. When  $\gamma \geq 2$  and the overtaken shock  $S_2^{(2)}$  is sufficiently weak, there exists a constant  $M_{0_1}^2$  (see (3.12)) such that

If  $M_0^2 < M_{0_1}^2$ , the reflected wave is the first kind of shock (Figure 3.16);

If  $M_0^2 > M_{0_1}^2$ , there exists unique value  $t_{M_0}^1$  such that

the reflected wave is the first kind of shock when  $1 \leq \frac{\rho_1}{\rho_0} < t_{M_0}^1$ ;

the reflected wave is the first kind of centered simple wave when

$t_{M_0}^1 < \frac{\rho_1}{\rho_0} \leq t^*$  (see Figure 3.17).

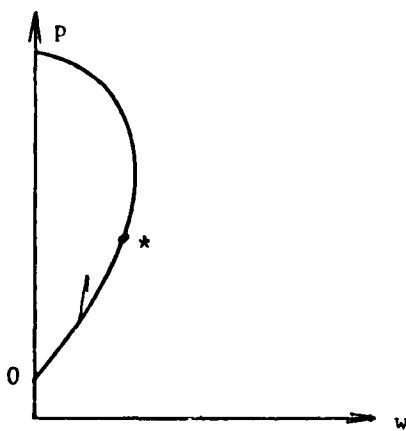


Figure 3.16

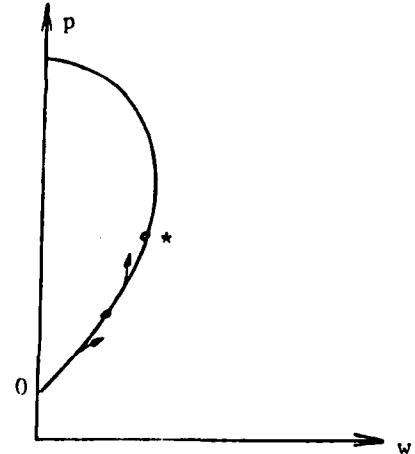


Figure 3.17

#### 4. The interaction of shock and contact discontinuity

It can be shown easily that the projection of  $S_2(0)$  into  $(p, w)$  plane is expressed by

$$w = \frac{\frac{p}{p_0} - 1}{\gamma M_0^2 - \frac{p}{p_0} + 1} \quad \frac{(1+\mu^2)(M_0^2-1) - (\frac{p}{p_0}-1)}{\frac{p}{p_0} + \mu^2} \quad (4.1)$$

here  $\mu^2 = \frac{\gamma-1}{\gamma+1}$ .

Let  $y = \frac{p}{p_0} - 1$  then  $0 \leq y \leq (1+\mu^2)(M_0^2-1)$  and (4.1) can be rewritten as

$$w = \frac{y}{\gamma M_0^2 - y} \quad \frac{(1+\mu^2)(M_0^2-1) - y}{y + (1+\mu^2)}.$$

Fix  $p_0$ , it is easy to see that the sign of  $\frac{\partial w}{\partial M_0^2}$  is the same as the

sign of  $L = L(y, M_0^2)$  where  $L = y + (2-M_0^2)$ . Therefore,  $\frac{\partial w}{\partial M_0^2} > 0$  when

$M_0^2 \leq 2$  (Figure 4.1) and  $\frac{\partial w}{\partial M_0^2} < 0$  when  $M_0^2 > 2$  and  $y = 0 + 0$

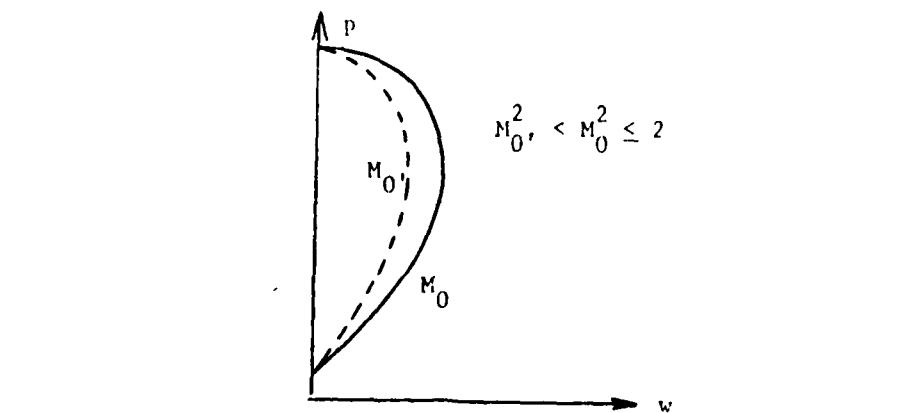


Figure 4.1

(Figure 4.2) which imply the following theorem.

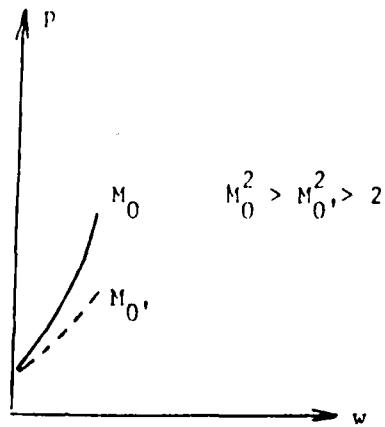


Figure 4.2

Theorem 4.1 Assume that a contact discontinuity  $T$  interacts with a shock  $S_2$ . Let the wave front state and wave back state of the shock  $S_2$  be (1) state and (2) state respectively. The states on the two sides of the contact discontinuity  $T$  are denoted by (0) state and (1) state respectively (Figure 4.3).

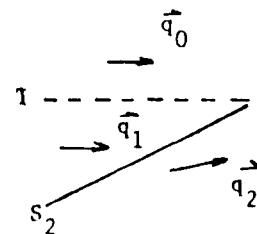


Figure 4.3

Suppose the magnitude of  $S_2$  is sufficiently weak, then it is certain that the shock  $S_2$  and the contact discontinuity  $T$  penetrate with each

P 34

other and a reflected wave comes out which is a centered simple wave

$R_1$  when  $M_0^2 < M_1^2 \leq 2$  or  $M_1^2 > M_0^2 > 2$  (Figure 4.4) and is a shock  $S_1$

when  $M_1^2 < M_0^2 \leq 2$  or  $M_0^2 > M_1^2 > 2$  (Figure 4.5). When 2 is between

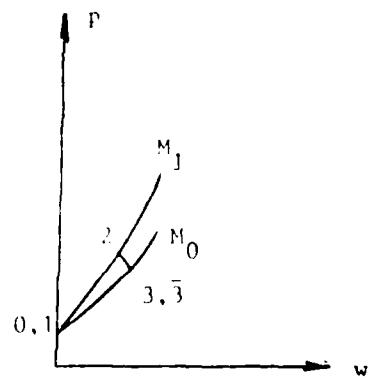
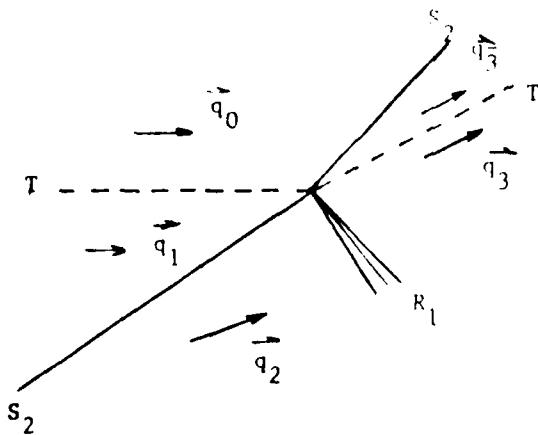


Figure 4.4

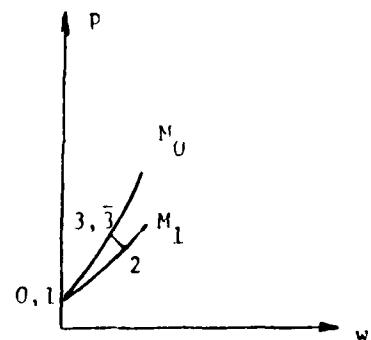
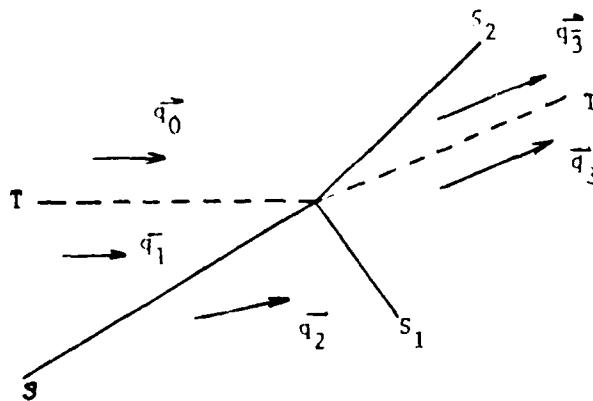


Figure 4.5

$M_1^2$  and  $M_0^2$ , no matter what the case is:  $M_0^2 > M_1^2$  or  $M_0^2 < M_1^2$ , the reflected wave may be either a centered simple wave  $R_1$  or a shock  $S_1$ .

## APPENDIX

The proof of (2.13).

Regarding  $p$  as the function of  $\rho$  in (2.7) and differentiating to  $\rho$ , it follows that

$$u \frac{dv}{d\rho} - v \frac{du}{d\rho} = \frac{1}{2\rho(v-\sigma u)^2} [\rho u] \begin{vmatrix} [\rho u] & v - \sigma u & -(v-\sigma u) \\ [\rho uv] & \frac{dp}{d\rho} + v(v-\sigma u) & -2v(v-\sigma u) \\ [p+\rho u^2] & u(v-\sigma u) - \sigma \frac{dp}{d\rho} & -2u(v-\sigma u) \end{vmatrix}$$

$$= \frac{1}{-2\rho(v-\sigma u)[\rho u]} \begin{vmatrix} [\rho u] & v - \sigma u & 1 \\ [\rho uv] - v[\rho u] & \frac{dp}{d\rho} & v \\ [p+\rho u^2] - u[\rho u] & -\sigma \frac{dp}{d\rho} & u \end{vmatrix}. \quad (5.1)$$

Due to  $[p+\rho u^2] - u[\rho u] = 0$  when  $v_0 = 0$ , it can be shown that

$$\Delta = [\rho u](u+\sigma v) \frac{dp}{d\rho} - \{[\rho uv] - v[\rho u]\} \{(v-\sigma u)u + \sigma \frac{dp}{d\rho}\}$$

where  $\Delta$  is the determinant in (5.1).

On account of  $[\rho uv] - v[\rho u] = u_0 \rho_0 v = \frac{c_0^2(\rho-\rho_0)}{b\sigma} = \frac{b(\rho-c_0)}{\sigma} \frac{dp}{d\rho}$ .  
 (By using  $\frac{dp}{d\rho} = \frac{c_0^2}{b}$ ;  $\frac{c_0^2}{b} = \frac{c_0^2}{b}$ ), it turns out

$$u \frac{dv}{d\rho} - v \frac{du}{d\rho} = \frac{1}{-2\rho(v-\sigma u)[\rho u]} \{[\rho u](u+\sigma v) - \frac{b(\rho-\rho_0)}{\sigma} (v-\sigma u)u - \frac{c_0^2}{b}(\rho-\rho_0)\}. \quad (5.2)$$

For simplicity, we restrict our attention to  $\sigma = \sigma_2$ . In view of the following:

$$u = u_0 - \frac{c_0^2}{u_0 \rho_0 b} (\rho - \rho_0) ;$$

$$[\rho u] = (u_0 + \rho \frac{u-u_0}{\rho-\rho_0}) (\rho - \rho_0)$$

$$= \frac{b'u_0^2 - c^2 \frac{\rho}{\rho_0}}{b'u_0} (\rho - \rho_0) ;$$

$$\rho(v - \sigma u) = \rho_0(v_0 - \sigma u_0)$$

$$= - \sigma u_0 \rho_0$$

$$= - \sqrt{\frac{c^2 \frac{\rho}{\rho_0}}{b'u_0^2 - c^2 \frac{\rho}{\rho_0}}} \cdot u_0 \rho_0 ;$$

$$u + v\sigma = u_0 .$$

We obtain that

$$[\rho u](u+v\sigma) = \frac{b(c-v_0)}{\sigma} (v-\sigma u) u - \frac{c_0^2}{b} (\rho - \rho_0)$$

$$= \frac{b'u_0^2 - c^2 \frac{\rho}{\rho_0}}{b'} (\rho - \rho_0) + \frac{u_0^2 \rho_0 b - c_0^2 (\rho - \rho_0)}{\rho} (\rho - \rho_0) - \frac{c_0^2 (\rho - \rho_0)}{b} .$$

And so

$$u \frac{dv}{dp} - v \frac{du}{dp} = \frac{(b'u_0^2 - c^2 \frac{\rho}{\rho_0})(\rho - \rho_0) + \frac{\rho_0 b}{\rho} \{ b'u_0^2 - \frac{b'c_0^2}{b\rho_0} (\rho - \rho_0) \} (\rho - \rho_0) - \frac{b'c_0^2 (\rho - \rho_0)}{b}}{2c_0(\rho - \rho_0) \cdot \sqrt{\frac{c^2 \frac{\rho}{\rho_0} (b'u_0^2 - c^2 \frac{\rho}{\rho_0})}{b'u_0^2 - c^2 \frac{\rho}{\rho_0}}}}$$

$$= \frac{\sqrt{\frac{c^2 \frac{\rho}{\rho_0} (b'u_0^2 - c^2 \frac{\rho}{\rho_0})}{b'u_0^2 - c^2 \frac{\rho}{\rho_0}}}}}{2c^2 \rho} \left\{ 1 + \frac{\rho_0 b}{\rho} - \frac{\gamma+1}{2} \frac{c^2}{b'u_0^2 - c^2 \frac{\rho}{\rho_0}} \frac{\rho - \rho_0}{\rho} \right\} .$$

By using  $\frac{c^2}{b^2} = \frac{c_0^2}{b}$  again, we have

$$\frac{dw}{dp} = \frac{\sqrt{c^2 b^2 p - (bu_0^2 - c_0^2 p_0)}}{2c^2 u^2 p} \quad \left[ 1 + \frac{v_0 b}{p} - \frac{1+1}{2} \cdot \frac{c_0^2}{bu_0^2 - c_0^2 p_0} \cdot \frac{p-p_0}{b} \right]$$

This proves (2.13).

REFERENCES

- [1] Courant, R. and Friedrichs, F.O., *Supersonic Flow and Shock Waves*, New York, 1965.
- [2] T. Zhang and Y.F. Guo, A special initial value problem for the system in gas dynamics, *Acta Mathematica Sinica*, 15(1965), 386-396.
- [3] L. Hsiao and T. Zhang, Interactions of Elementary Waves in One-dimensional Adiabatic Flow, *Acta Mathematica Sinica*, 22(1979), 596-619.

HOMOGENIZATION, CONVEX ANALYSIS, AND THE GEOMETRY  
OPTIMIZATION OF ENGINEERING STRUCTURES

by

ROBERT V. KOHN  
Courant Institute of Mathematical Sciences

Gilbert Strang  
Massachusetts Institute of Technology

Lecture by Gilbert Strang.

## I. HOMOGENIZATION, GENERALIZED STRUCTURES, AND OPTIMIZATION.

In many different areas of structural design engineering, one wishes to consider isoperimetric problems of the following type: how should one choose the shape of a structure so as to minimize its cost subject to constraints on its strength?

Questions of this type have been studied at great length by both mathematicians and structural engineers, in a wide variety of contexts. From a practical point of view truly optimal structures are often unwise designs, being highly unstable and potentially very weak with respect to loads other than those for which they were designed. Nonetheless, it is of obvious value to know how the optimal structure looks and what it costs, in order to approach its performance with more practical designs if possible.

To fix ideas we begin by describing a typical problem in this class. Consider a homogeneous, linearly elastic material characterized by a stress-strain law

$$(1) \quad \sigma_{ij} = \sum_{k,l=1}^3 A_{ijkl} \epsilon_{kl}$$

where  $\epsilon_{kl}(u) = \frac{1}{2}(\partial u^k / \partial x^l + \partial u^l / \partial x^k)$  is the linearized strain of a displacement  $u$ . Consider a smoothly bounded region  $\Omega \subset \mathbb{R}^3$ , with  $\partial\Omega$  decomposed as  $\Gamma_0 \cup \Gamma_1$ ; we take  $\Gamma_0$  to be clamped and  $\Gamma_1$  to be loaded by a fixed force  $f: \Gamma_1 \rightarrow \mathbb{R}^3$ . Given  $0 < V < \text{Vol}(\Omega)$ , we define the class of "admissible structures"

$$\mathcal{J}(v) = \left\{ S \subset \Omega : \partial S \text{ is Lipschitzian, } \text{Vol}(S) = v, \text{ and } \mathcal{L} \subset \partial S \right\}.$$

Each such structure  $S$  responds to the load  $f$  by a deformation  $u_S$ , determined by solving the elliptic system

$$(2) \quad \begin{aligned} \sum_{j=1}^3 \frac{\partial}{\partial x_j} \sigma_{ij} &= 0 \quad \text{on } S, \quad i=1,2,3 \\ \sigma_{ij} n_j &= \begin{cases} f & \text{on } \Gamma_1 \\ 0 & \text{on } \partial S \sim \partial \Omega \end{cases} \\ u_S &= 0 \quad \text{on } \Gamma_0, \end{aligned}$$

where  $n$  denotes the unit normal vector to  $\partial S$  and  $\sigma$  depends on  $u_S$  by (1).

We identify the compliance of the structure — the work done by the load  $f$  — as

$$c(u_S, f) = \int_{\Gamma} u_S \cdot f \quad .$$

A typical geometry optimization problem, then, is

$$(3) \quad \text{Minimize} \left\{ c(u_S, f) : S \subset \mathcal{J}(v) \right\} \quad .$$

In words: one wants to remove a given volume from  $\Omega$  such that what is left has the minimum possible compliance under the load  $f$ .

This example represents perhaps the simplest fully three dimensional case of geometry optimization. Many, many variations occur in situations of practical interest. We have taken as "cost" the volume of material; this is rather typical — particularly in problems of aerospace design,

where weight is at a premium. One might, however, have several different materials at one's disposal, each with a different cost per unit volume and with different material properties. While the compliance is a mathematically convenient notion of strength, one might wish to control the pointwise supremum of the stresses instead; for ductile materials the limit multiplier of the load is an appropriate parameter; or one could try to use nonlinear elasticity and modern theories of fracture. In buckling or vibration problems the quantity of interest is often the lowest eigenvalue of an associated elliptic equation. Instead of considering a single, fixed load, one might want to optimize the structure's performance under several loads at once, or under a random distribution of loads. And in addition to fully three dimensional structures, such problems arise naturally for axially symmetric rods in torsion, flat plates in plane strain, variable-thickness plates in bending, curved shells, etc.

It's easy to get carried away formulating problems, however, and quite another thing to solve them. Not surprisingly, almost all progress has been restricted to the linearized models of behavior: linear elasticity and linearly elastic-perfectly plastic materials. The literature in these areas is vast, and a comprehensive review of it is far beyond the scope of this paper — the interested reader may refer to [1] and [2] for review articles and further references. What we propose to do here is to summarize — in a highly selective and idiosyncratic fashion — some of the major ideas in the field.

A great deal of attention has been directed toward sensitivity analysis and the development of gradient flow techniques.[3,4,5,6] This work views our

optimization problems as special cases of the optimal control theory of distributed parameter systems, in which the control variable is the domain on which a given partial differential equation is to be solved. These methods apply primarily to problems where one is solving an elliptic equation (elasticity, but not plastic limit analysis), and where the strength is an integral functional of the solution (compliance, not maximum stress).

The most elementary product of this approach is a necessary condition for optimality, obtained by taking the "first variation" of the optimal domain. For example, if a smoothly bounded set  $S_0$  is optimal for problem (1), then its associated displacement  $u_0$  must satisfy [1]

$$(4) \quad \begin{aligned} \|\varepsilon(u_0)\|^2 &\geq c && \text{on } S_0 \\ \|\varepsilon(u_0)\|^2 &= c && \text{on } \partial S_0 \end{aligned}$$

for some constant  $c > 0$ , where we denote by  $\|\varepsilon(u)\|^2$  the associated energy per unit volume

$$\|\varepsilon(u)\|^2 = \sum_{i,j,k,l} A_{ijkl} \varepsilon_{ij}(u) \varepsilon_{kl}(u) .$$

In fact, for compliance problems (and for plastic limit multiplier problems) one can give a sufficient condition for optimality that is closely related to (4). An entirely elementary argument -- using only the fact that (1) is equivalent to a certain variational problem -- shows that if, for some  $S \subset \mathcal{F}(V)$  and  $c > 0$  the deformation  $u_S$  extends to an element of  $H^1(\Omega; \mathbb{R}^3)$  satisfying

$$(S) \quad \begin{aligned} \|\Sigma(u_S)\|^2 &\geq c \quad \text{a.e. on } S \\ \|\Sigma(u_S)\|^2 &\leq c \quad \text{a.e. on } \Omega - S \end{aligned}$$

[ $\square$ ]  
then  $S$  is optimal for ( $\Xi$ ). While the computation leading to ( $\Psi$ ) is very general, the sufficiency of ( $S$ ) rests upon the special relationship between the compliance of  $S$  and the variational form of ( $\Xi$ ).

Given a structure  $S$  that is not optimal, the same computation that yields ( $\Psi$ ) gives the gradient of the compliance in the space of local deformations of  $S$ . This leads to the formulation of gradient flow algorithms for finding structures that are at least local optima. Although the computations are usually too onerous to be practical, some use has been made of this method [8, 9]. In addition, a "fixed point method" for satisfying the optimality conditions ( $\square$ ) directly has been used for model problems of compliance with good preliminary results [ $\square$ ]. None of these algorithms has been shown to converge, however; indeed, it has been unclear whether to expect a smoothly bounded optimal set  $S$  to exist at all for a problem such as ( $\Xi$ )!

In a variety of special problems, optimal structures have been shown to exist, and in some cases they can even be given explicitly. For certain two-dimensional problems, complex variable methods can be applied [10, 12]. In other problems — principally the case of plastic rods in torsion — specific formulas for the "strength" allow one to identify the optimal structure. [13, 14] And in yet other cases, symmetrization has been used [15]. It must be said, however, that many of these methods have a somewhat ad hoc flavor; they represent, one senses, something less than

a general picture of what optimal geometries can look like.

In fact, there have been indications that in many cases optimal structures will exist only in a generalized sense. In other problems of distributed parameter control such a phenomenon was noticed by Murat [6]. In torsion problems this phenomenon was noticed by Lurie and Kłosowicz [17]. In fact it is well-known folklore within the structural design optimization community that as one tries to optimize a three dimensional structure — for example, in our problem (3) — the structure may develop many small holes in such a manner as to mimic the behavior of an optimal "truss-like continuum"; numerical experimentation along these lines may be found, for example, in [8].

That optimal shapes might fail to exist should be no surprise to a mathematician familiar with recent work concerning "homogenization of domains" [19]. There one finds that if a sequence  $S_n$  is defined by perforating a domain  $\Omega$  with holes of a fixed geometry but rescaled to a lattice of size  $1/n$ , then the corresponding displacements  $u_n$  converge in a suitable weak sense to the solution of a new equation, now defined on all of  $\Omega$ , of the same type as (2) but with a stress-strain law that depends upon the local geometry of the holes — and which is, in this case, explicitly computable. In short: new, "effective materials" may be produced from the original one by allowing geometric microstructures to develop. One says that the "effective materials", or the equations characterizing them, are obtained by homogenization (also known as  $\Gamma$ -convergence) from the original equations.

There is, to be sure, nothing in our original problem that requires the local geometry of the microstructure to be periodic. To say something mathematically rigorous about the general situation, however, one must modify the problem slightly. Given a set  $S \in \mathcal{J}(V)$ , let us not leave  $\Omega \setminus S$  empty, but fill it instead with a second (perhaps very weak) homogeneous linearly elastic material, whose stress-strain law is

$$\sigma'_{ij} = \sum_{k,l=1}^3 \tilde{A}_{ijkl} \epsilon_{kl}$$

The displacements must now satisfy the equilibrium equations

$$(6) \quad \begin{aligned} \sum_{j=1}^3 \frac{\partial}{\partial x_j} \sigma'_{ij} &= 0 \quad \text{on } \Omega \quad (i=1,2,3) \\ \sigma \cdot n &= f \quad \text{on } \Gamma_1 \\ u_S &= 0 \quad \text{on } \Gamma_0 \end{aligned}$$

where

$$\begin{aligned} \sigma'_{ij} &= \sum_{k,l=1}^3 A_{ijkl} \epsilon_{kl}(u_S) \quad \text{on } S \\ &= \sum_{k,l=1}^3 \tilde{A}_{ijkl} \epsilon_{kl}(u_S) \quad \text{on } \Omega \setminus S. \end{aligned}$$

The definition of the compliance  $c(u_S, f)$  remains unchanged, as does the form of the optimization problem:

$$(7) \quad \text{Minimize } \{c(u_S, f) : S \in \mathcal{J}(V), \text{ and } u_S \text{ solves (6)}\}.$$

Because the system (6) is elliptic on all of  $\Omega$ , we can invoke a compactness theorem due essentially to Spagnolo [20, 21] to conclude the existence of effective materials in general. Given any sequence  $\{S_n\} \subset \mathcal{J}(V)$ ,

there is a subsequence  $\{S_{n(j)}\}$  with the following property: for each  $f \in H^{-1}(\Omega; \mathbb{R}^3)$  the solutions of (7) converge weakly in  $H^1(\Omega; \mathbb{R}^3)$  to a solution  $\bar{u}$  of equations of the same type as (6), with a new stress-strain law

$$(8) \quad \sigma_{ij} = \sum_{k,l=1}^3 A_{ijkl}^{\text{eff}}(x) \varepsilon_{kl}$$

corresponding in general to an inhomogeneous, anisotropic, linearly elastic "effective material". The new stress-strain law (8) depends only on the subsequence  $\{S_{n(j)}\}$ , not on the load  $f$ . Moreover, the compliances converge:

$$c(u_{S_{n(j)}}, f) \rightarrow c(\bar{u}, f)$$

for each  $f$ .

Thus the existence of solutions to (7) becomes a triviality if one allows as generalized solutions the effective materials that arise by homogenization in the manner just described. From this point of view, the interesting problems are these:

(9) 1) What are the effective equations that can be produced using sets in  $\mathcal{J}(V)$ ?  
 2) What do the optimal "generalized solutions" look like?

Question (9-1) has considerable interest over and above its relevance to optimization problems. Given an answer to it, one should be able to handle (9-2) successfully by means of first-order optimality conditions and gradient-flow methods.

Unfortunately, answering (9-1) seems to be a difficult task. Some limited progress has been made: for Laplace's equation in  $\mathbb{R}^2$ , Tartar

has characterized the limiting equations obtainable by homogenizing two isotropic constant coefficient ones without regard to volume fraction used [1, 2]. In more general situations, or when trying to take volume fractions into account, one can say much less — in general, only rather crude bounds are available [22, 24, 25–28]. Settling this question remains an important, open problem in the theory of homogenization.

We have so far touched only briefly upon the engineering literature. As indicated earlier, the idea that some sort of "generalized structure" will be required to describe optimal solutions is by no means new to the engineering community. Rather than belabor the question of what generalized structures can be made from given materials, however, most of their work passes directly to consideration of the generalized structures themselves. In many contexts this amounts simply to enlarging the class of materials one is willing to work with, so that a continuous range of materials is available, each with a preassigned cost; in other cases one allows structures of an entirely new class. The cases that have received the most attention are "truss-like continua" (for three-dimensional problems and two-dimensional plane strain), "grillage-like continua" (for planar structures supporting bending loads), and variable-thickness plates. [29, 30]. In most cases the relevant optimization problems are formulated in finite-dimensional versions, with the continuous version obtained by a formal passage to the limit. In many cases involving the compliance of an elastic structure or the limit multiplier of a perfectly plastic one, one can use the theory of convex duality to great advantage.

Perhaps we can give some flavor of this approach by describing the

analogue of problem (1) in the category of truss-like continua. Such a structure is described by a finite family of vector fields, say  $\{\tau^j\}_{j=1}^N$ ; here  $\tau^j/|\tau^j|$  determines the direction of the  $j$ 'th family of truss members, and  $|\tau^j|^2$  their strength per unit length, which we identify also as the cost per unit length. We assume that  $|\tau^j|^2$  can be arbitrarily large, that joining members of the "truss" can be done at no cost, and that one can ignore the possible buckling of truss members.

For such a structure, the analogue of the equilibrium equations (1) is most easily expressed in variational form:

$$(10) \quad \min_{u=0 \text{ on } \Gamma_0} \frac{1}{2} \int_0^L \sum_{j=1}^N \langle \Sigma(u), \tau^j \otimes \tau^j \rangle^2 dx = \int_{\Gamma} f \cdot u .$$

The design optimization problem is

$$(11) \quad \text{Minimize } \left\{ \int_{\Gamma} f \cdot u : u \text{ solves (10) with } \int_{\Omega} \sum_{j=1}^N |\tau^j|^2 dx \leq c \right\}.$$

We have not identified appropriate spaces for either  $\tau^j$  or  $u$ ; indeed, it is not clear what choices one should make, and the above formulas should be considered formal only.

One can characterize solutions to (11) — once again, on a purely formal level — by means of convex duality. One is led to conclude that in an optimal structure  $N=3$ , and that the solution  $u$  of (10) has eigenvectors  $\tau^j/|\tau^j|$  with eigenvalues  $\pm c$ , for some constant  $c$ , whenever  $\tau^j \neq 0$ .

Many interesting mathematical questions remain open here. A correct mathematical treatment of truss-like continua has yet to be given, as does a proof of the existence of optimal structures in this class. It would be useful to have a regularity theory and methods for computing optima as well. There are other applications of convex duality or the

Kuhn-Tucker conditions (in engineering optimization these are usually called the Prager-Shield conditions) that have yet to be carried out correctly in infinite-dimensional contexts; this is particularly interesting for plasticity problems, where the spaces one must work in are rather unfamiliar [28].

Finally, we emphasize that whatever class of objects one takes as admissible structures, one must still consider the possibility that unexpected, generalized structures can be produced by a limiting process. Recent work by Cheng and Olhoff [30] has found this to occur, for example, in a previously unexpected manner in the optimal design of variable-thickness elastic plates.

### III. OPTIMAL CROSS SECTIONS FOR RODS IN ANTIPLANE SHEAR

In this section we summarize some recent work concerning the optimal geometry of the cross section of a rod loaded in antiplane shear. The details of this work will appear soon elsewhere [31]; our goal here is to describe the methods used, which seem rather general and potentially applicable far beyond the context of the model problem discussed here.

We consider rods of infinite length and constant cross section, loaded by a boundary shear force directed along the length of the rod, uniformly along that length. As the strength criterion we take the plastic limit multiplier of the load, though we will comment on the corresponding

compliance problem at the end of the section.

The geometry optimization problem is once again cast in terms of removal of material: how should a fixed amount of area be removed from the interior of a rod cross section so as to weaken the structure as little as possible? The key to our approach is that we neither attempt to characterize all homogenized, generalized structures that might occur, nor do we merely assume the properties of some specific generalized structure. Rather, we characterize those microstructures that arise in optimal configurations; in other words, we derive the correct class of generalized structures. One can then apply infinite-dimensional convex analysis in a manner parallel to that used in the engineering literature, to characterize the optimal structures.

So let  $U \subset \mathbb{R}^2$  represent the rod's section before volume removal, and assume  $\bar{\Gamma} = \partial U$  is piecewise smooth. The load  $f: \bar{\Gamma} \rightarrow \mathbb{R}$  should be bounded and measurable, and since (for simplicity only) none of  $\bar{\Gamma}$  is clamped one has a consistency condition  $\int f = 0$ . The geometry of the model problem is represented in figure 1.

Consider the class of admissible cross sections  $\mathcal{U}$ , defined by

$$\mathcal{U} = \left\{ U' \subset U : \partial U' \text{ is Lipschitzian, and } \bar{\Gamma} \subset \partial U' \right\}.$$

We define, for  $U' \in \mathcal{U}$ ,

(12)  $U'$  withstands load  $f$  iff There exists  $\sigma \in L^\infty(U'; \mathbb{R}^2)$  such that  
 $\operatorname{div} \sigma = 0$  on  $U'$ ;  $\sigma \cdot n = f$  on  $\bar{\Gamma}$ ;  
 $\sigma \cdot n = 0$  on  $\partial U' \setminus \bar{\Gamma}$ ; and  $|\sigma| \leq 1$  a.e.

This definition corresponds to the model of plastic limit analysis, with "yield condition"  $|\sigma| \leq 1$ . The vector  $\sigma$  represents the shear stresses in the rod (all other stresses are zero). In words,  $U'$  withstands the load  $f$  if there is some stress which equilibrates the load  $f$  and which nowhere exceeds the yield condition. Of course, since  $\sigma$  is merely an  $L^\infty$  vector field the condition  $\operatorname{div} \sigma = 0$  must be understood weakly; also, we are using the fact that  $\sigma \cdot n$  has an  $L^\infty$  trace on  $\partial U'$ . Finally, we remark that the unknown boundary  $\partial U' \cap \Gamma'$  is always unloaded and unclamped.

Let us touch base with the more familiar terminology of plasticity theory. The limit multiplier  $\lambda(U', f)$  is defined as

$$\lambda(U', f) = \sup \left\{ t : U' \text{ withstands load } tf \right\}.$$

The duality theory of plastic limit analysis provides a useful tool for determining whether or not  $U'$  withstands load  $f$  [32]:

$$(13) \quad \lambda(U', f) = \inf \left\{ \int_{U'} |\nabla u| : u \in H^1(U'; \mathbb{R}), \int_{\Gamma'} u \cdot f = 1 \right\}.$$

One can further interpret (13) in terms of a very geometric isoperimetric-type problem in the plane, which allows one to solve for  $\lambda(U', f)$  in many cases.

We come to an elementary, but crucial, observation: if one extends the vector field  $\sigma$  in (12) to the "hole"  $U \cup U'$  by assigning it the value zero, this extension remains divergence-free. Thus

$U'$  withstands load  $f$  iff There exists  $\sigma \in L^\infty(U; \mathbb{R}^2)$  such that  
 $\operatorname{div} \sigma = 0$  on  $U$ ;  $\sigma \cdot n = f$  on  $\Gamma'$ ;  
 $\sigma = 0$  a.e. on  $U \cup U'$ ; and  $|\sigma| \leq 1$  a.e.

For the purposes of our theory, it is convenient to recast the geometry optimization problem slightly, fixing not the amount of area to be removed but instead the strength of the result. In this form, the problem is: given  $0 < t \leq \lambda(U, f)$ , find

$$(17) \quad P_t = \inf \left\{ \text{Area}(U') : U' \in \mathcal{U}, \lambda(U', f) \geq t \right\}$$

and describe an optimizing sequence of sets  $U'$ .

The key to solving this problem is the following lemma.

Lemma 1: Let  $\sigma \in L^\infty(U; \mathbb{R}^2)$  with  $\text{div } \sigma = 0$  and  $|\sigma| \leq 1$ . For each  $\epsilon > 0$  one can construct a set  $U_\epsilon \in \mathcal{U}$  and a vector field  $\tilde{\sigma}_\epsilon \in L^\infty(U; \mathbb{R}^2)$  such that

- i)  $\text{Area}(U_\epsilon) \leq \int_U \sigma_1 + \epsilon$
- ii)  $\tilde{\sigma}_\epsilon = 0$  on  $U \setminus U_\epsilon$
- iii)  $|\tilde{\sigma}_\epsilon| \leq 1$ ,  $\text{div } \tilde{\sigma}_\epsilon = 0$ , and  $\tilde{\sigma}_\epsilon \cdot n = \sigma \cdot n$  on  $\partial U$ .

The proof of lemma 1 is rather technical, but the idea is simple: one replaces a region where  $0 < |\sigma| < 1$  by a foliation of slits parallel to  $\sigma$ , leaving behind density  $|\sigma|$  of material; choose  $\tilde{\sigma}_\epsilon$  in such a region to be parallel to the slits with  $|\tilde{\sigma}_\epsilon| = 1$ , except of course on the slits where  $\tilde{\sigma}_\epsilon = 0$ . There is, to be sure, some work to be done to show that this can be done even if  $\sigma$  is in no way smooth, and that these local pictures can be pieced together; the complete argument will appear in [3].

One may rewrite (17) heuristically as

$$P_t = \inf \left\{ \int_U \chi_{\sigma \neq 0} : \sigma \in L^\infty(U; \mathbb{R}^2), |\sigma| \leq 1, \text{div } \sigma = 0, \sigma \cdot n = tf \text{ on } \Gamma \right\},$$

where  $\chi_{\sigma \neq 0}$  represents the characteristic function of the set where  $\sigma \neq 0$ . (This is only heuristic, because the set where  $\sigma \neq 0$  may or may not be regular enough to belong to the admissible class  $\mathcal{U}$ .)

Using Lemma 1, one readily sees that in fact

$$(15) \quad P_t = \inf \left\{ \int_U |\sigma| : \sigma \in L^\infty(U; \mathbb{R}^2), |\sigma| \leq 1, \operatorname{div} \sigma = 0, \sigma \cdot n = tf \text{ on } \Gamma \right\}.$$

Thus on the heuristic level the role of lemma 1 is to identify the integrand  $\int_U |\sigma|$  as the lowersemicontinuous hull of  $\int_{\sigma \geq 0}$ ! Given  $\sigma$  solving (15), one can construct an optimizing sequence of shapes by the construction implicit in lemma 1.

The integrand  $\int_U |\sigma|$  is convex — this, it seems, is a fortunate coincidence; one had no right to expect it to be, a priori. However, since it is convex one can achieve additional insight concerning (15) by considering simultaneously its convex dual. Moreover, the existence of an optimal  $\sigma$  solving (15) is immediate from the weak\* compactness of the unit ball in  $L^\infty$ . Applying duality theory leads to the following result [33].

Theorem 1:

A) The infimum in (15) is attained by an  $L^\infty$  vector field.

B) The optimal area  $P_t$  is also the value achieved by the dual problem

$$(16) \quad P_t = \sup \left\{ \int_U (1 - |\nabla u|)_+ + t \int_U u \cdot f : u \in H^1(U) \right\}$$

and the supremum is attained if we allow  $u$  to lie in the larger space  $BV(U)$ .

C) If  $\sigma_t \in L^\infty(U; \mathbb{R}^2)$  and  $u_t \in BV(U)$  are solutions of (15) and (16) respectively, they satisfy the saddle-point condition

$$(17) \quad \int_U |\sigma_t| - \langle \sigma_t, \nabla u_t \rangle = \int_U (1 - |\nabla u_t|)_+$$

which implies in particular

$$i) \quad \sigma_t = \frac{\nabla u_t}{|\nabla u_t|} \quad \text{a.e. where } |\nabla u_t|_{\text{abs}} > 1$$

$$ii) \quad \sigma_t / |\sigma_t| = \nabla u_t \quad \text{a.e. where } 0 < |\sigma_t| < 1$$

$$iii) \quad \sigma_t = 0 \quad \text{a.e. where } |\nabla u_t|_{\text{abs}} < 1.$$

D) If  $\sigma \in L^\infty(U; \mathbb{R}^2)$  with  $|\sigma| \leq 1$ ,  $\operatorname{div} \sigma = 0$ ,  $\sigma \cdot n = t_f$  on  $\Gamma$ ;  $u \in BV(U)$ ; and if (7) holds for  $u$  and  $\sigma$  then  $u$  and  $\sigma$  are extremal for (1') and (1'') respectively.

Many of these statements must be understood in a rather weak sense, since we have asserted very little regularity for  $u$  and  $\sigma$ . By  $(1-|\nabla u|)$  one understands  $(1-|\nabla u|)_{\text{abs}}$  dx -  $|\nabla u|_{\text{sing}}$ , where  $|\nabla u| = |\nabla u|_{\text{abs}} + |\nabla u|_{\text{sing}}$  is the decomposition of the measure  $|\nabla u|$  into its absolutely continuous and singular parts, and for real numbers  $p$ ,  $p_- = \min\{p, 0\}$ . One must verify that when  $\operatorname{div} \sigma = 0$  and  $|\sigma| \leq 1$ , the integral  $\int \langle \sigma, \nabla u \rangle$  makes sense for each  $u \in BV(U)$ . And implicit in (C), (i-iii) is the fact that for  $u \in BV(U)$ , the unit vector  $\nabla u / |\nabla u|$  is well-defined  $|\nabla u|$ -almost everywhere.

We expect to be able to prove further regularity for the extremals to (1') and (1''): it appears that  $u_t$  is locally Lipschitzian, and that  $\sigma_t$  is a  $C^1$  vector field away from the set  $\{x: \sigma_t(x) = 0\}$  — which itself has piecewise  $C^1$  boundary. As of this writing, however, some details remain before the proof of these assertions can be considered complete.

Theorem 1 clarifies greatly — at least in the model of antiplane shear — the role played by optimality conditions such as we discussed in section I. The analogue in our model problem of the conditions (8)

is this : if  $U' \in \mathcal{U}$ ,  $u' : U \rightarrow \mathbb{R}$  achieves  $\inf \left\{ \int_U |\nabla u| : \int_U u \cdot f = 1 \right\}$ ,  
and for some  $c > 0$

$$|\nabla u'| \geq c \quad \text{on } U'$$

$$|\nabla u'| \leq c \quad \text{on } U \sim U'$$

then  $U'$  has minimal area for its limit multiplier. The prospect of finding such a set  $U'$  may seem bleak, since in general a minimizer for (13) may well have jump discontinuities. However, Theorem 1 shows that this criterion comes close to the mark: an optimal set  $U'$  may not exist, in general, but in a certain sense  $u'$  does nonetheless — it may be identified, up to scaling, with the function  $u_t$  in Theorem 1. The optimal stress  $\sigma_t$  is the closest thing to an optimal geometry that exists in general; where  $\sigma_t = 0$  one has a hole; where  $\sigma_t = 1$  one leaves the original section unchanged, and where  $0 < \sigma_t < 1$  one obtains "near-optimal" structures by removing slits parallel to  $\sigma_t$ , as discussed earlier.

The saddle point conditions (17) enable one to construct examples with relative ease. An instructive case is that of a "butterfly-shaped" section, loaded uniformly along the "wing-tips" only (figure 2). The optimal stress field for  $t=1$  is shown in figure 3. Where the wings join the body the integral curves of  $\sigma$  are arcs of circles, and  $|\sigma|=1$ ; in the darkened regions  $\sigma=0$ ; and elsewhere the integral curves of  $\sigma$  are straight line segments. Figure 4 shows the corresponding function  $u$ , by marking some of its level curves. Where  $|\sigma|=1$ ,  $|\nabla u| \geq 1$  and the level curves are straight line segments; where  $0 < \sigma < 1$ ,  $|\nabla u| = 1$ ; in this case  $|\nabla u|=1$  where  $\sigma=0$ . We remark that if  $u$  is  $C^2$  and  $|\nabla u|=1$  on an open set then quite generally the integral curves of  $\nabla u / |\nabla u|$  must be

straight line segments.

Now consider what happens if the parameter  $t$  tends to zero. This corresponds physically to removing essentially all the material from the cross section. Formally, one might expect  $\sigma_t^* = t^{-1} \sigma_t$  to converge to a solution of

$$(18) \quad \inf \left\{ \int_U |\sigma| : \sigma \in L^1(U; \mathbb{R}^2), \operatorname{div} \sigma = 0, \sigma \cdot n = f \right\}$$

since each  $\sigma_t^*$  solves the corresponding problem with the added constraint  $|\sigma_t^*| \leq t^{-1}$ . The unit ball of  $L^1$  is not compact under weak convergence, however, so one should expect extremals for (?) to lie in a larger space. The correct place to look is the space of one-dimensional normal currents, as developed in [24]. Roughly speaking, these are vector valued measures that can be approximated by  $C^1$  vector fields, viewed as elements of the dual space to the one-dimensional differential forms. We have proven the following.

Theorem 2: Let  $\sigma_t$  and  $u_t$  be solutions of (7) and (8) respectively, and  $\sigma_t^* = t^{-1} \sigma_t$ . For any sequence  $t_n \rightarrow 0$ ,

A)  $\{\sigma_{t_n}^*\}$  has a subsequence which converges weakly to a normal current which is extremal for (?);

B)  $\{u_{t_n}\}$  has a subsequence which converges weakly in  $L^1(U)$  to a Lipschitzian function solving the <sup>problem that u</sup> dual to (?);

$$(19) \quad \sup \left\{ \int_U u \cdot f ; u: U \rightarrow \mathbb{R} \text{ Lipschitzian with } |u| \leq 1 \right\}$$

C) the extremal values of (18) and (19) are the same.

Figure 5 shows the "integral curves" of the solution to (?) for our butterfly example. The heavy segments at the top and bottom of the "body" carry positive mass of this current.

The methods sketched herein extend readily to a number of related problems. We conclude this section by indicating what implications the method has for linearly elastic rods under antiplane shear, with the "strength" interpreted in terms of the compliance.

The geometry of the problem remains the same as before, though now the load  $f$  may lie in  $H^{-\frac{1}{2}}(\Gamma)$ . The vertical displacement solves the equation

$$(17) \quad \Delta \bar{u} = 0 \text{ on } U', \quad \partial_n \bar{u} = f \text{ on } \Gamma, \quad \partial_n \bar{u} = 0 \text{ on } \partial U' \cup \Gamma;$$

equivalently,  $\bar{u}$  achieves the extremum in

$$(18) \quad \inf \left\{ \frac{1}{2} \int_{U'} |\nabla u|^2 - \int_{\Gamma} f \cdot u : u \in H^1(U') \right\}.$$

Integrating by parts in (17), we note that

$$\int_{U'} |\nabla \bar{u}|^2 - \int_{\Gamma} f \cdot \bar{u} = c(U', f).$$

As in the case of plastic limit analysis it is convenient to deal with the stresses rather than the displacements. In this problem the stress  $\bar{\sigma} = \nabla \bar{u}$  solves the dual to (18):

$$\inf \left\{ \frac{1}{2} \int_{U'} |\sigma|^2 : \sigma \in L^2(U'; \mathbb{R}^2), \operatorname{div} \sigma = 0, \sigma \cdot n = f \text{ on } \Gamma, \sigma \cdot n = 0 \text{ on } \partial U' \cup \Gamma \right\}.$$

Thus a geometry  $U'$  has compliance at most  $C$  if and only if there exists  $\sigma \in L^2(U'; \mathbb{R}^2)$  such that  $\operatorname{div} \sigma = 0$  on  $U'$ ,  $\sigma \cdot n = f$  on  $\Gamma$ ,  $\sigma = 0$  on  $U \sim U'$ , and  $\frac{1}{2} \int_{U'} |\sigma|^2 \leq C$ . Using this observation and the method of Lemma 1 we have proven the following.

Theorem 3. For each  $C > 0$ , the following quantities are equal :

- i)  $\inf \{ \text{Area}(U') : U' \in \mathcal{U}, c(U', f) \leq C \}$
- ii)  $\sup_{\alpha > 0} \inf \left\{ \int_{U'} F_\alpha(|\sigma|) dx - \alpha C : \sigma \in L^2, \operatorname{div} \sigma = 0, \sigma \cdot n = f \right\}$ .

$$\text{where } F_\alpha(t) = \begin{cases} 1 + (\alpha/2) t^2 & t \geq (2/\alpha)^{\frac{1}{2}} \\ (2\alpha)^{\frac{1}{2}} t & t \leq (2/\alpha)^{\frac{1}{2}} \end{cases}$$

$$\text{iii) } \sup_{\alpha > 0} \sup_{u \in H^1} \int_U \left( 1 - \frac{1}{2\alpha} |\nabla u|^2 \right)_+ + \int_U u f = \infty .$$

Again, details and examples are in preparation for publication elsewhere soon.

### III. DIRECTIONS FOR THE FUTURE.

It should be apparent that there is a great deal yet to be understood in the area of geometry optimization. We list here some directions that future work is likely to take.

- 1) Even for the model case of antiplane shear discussed in section II one can not yet characterize geometries that are optimal for their performance under several loads at once. Ultimately, one would like to study performance under random loads as well.
- 2) There should be an analogue of Lemma 1 for plane stress or for three dimensional problems. What integrand replaces  $\int_M f$ ? (We believe that for these problems the relevant integrand may not be convex.)
- 3) Can one do a similar analysis for eigenvalue problems?
- 4) The one-dimensional currents discussed in the context of theorem 2 play a role for antiplane shear analogous to that of truss-like continua in plane stress or three dimensional elasticity. It should be possible to

study the optimization of truss-like continua using methods from geometric measure theory. Particularly welcome would be algorithms for computing such optima.

- 5) Many biological structures have well-defined microstructures, which one expects serve to optimize their performance for certain tasks; certain bone and muscle tissues are striking cases of this. It would be interesting to "explain" such geometries by identifying with precision functions for which specific structures are designed optimally. Qualitatively this is an idea long familiar to biological scientists; but in rather few cases has it been made quantitative.
- 6) Any new result in the existence or characterization of optimal structures should lead to new methods for computing them and to convergence results for various algorithms. In particular, one should be able to use our Theorem 3 to study the "fixed point method" proposed for compliance-type problems in [24].
- 7) Perhaps even the general question of characterizing effective materials may yield to similar methods, by considering sufficiently complicated optimization problems.

## REFERENCES

- 1) Sawczuk, A. and Mroz, Z. Optimization in Structural Design, Proceedings of the IUTAM Conference at Warsaw, Springer-Verlag, 1975
- 2) Haug, E. and Cea, J.; Proceedings of NATO-ASI on Structural Optimization and Distributed Parameter Systems, Iowa City, 1980. Sijthoff and Nordhoff, to appear.
- 3) Cea, J., Identification of Domains, in Lecture Notes in Computer Science Vol. 3, Springer Verlag, 1973
- 4) Cea, J.; Gioan, A.; Michel, J.: Adaptation de la Methode du Gradient a un Probleme d'Identification de Domaine, in Lecture Notes in Computer Science, Vol 11, Springer-Verlag, 1974.
- 5) Dervieux, A., and Palmerio, B., Une Formule de Hadamard dans les Problemes d'Optimal Design, Lecture Notes in Computer Science, Vol. 40, Springer - Verlag, 1976.
- 6) Zolesio, J.P. Identification de Domaine par Deformation, These, Universite de Nice, 1979.
- 7) Mroz, Z., Archiwum Mech. Stosow. 15, pg. 63 (1963).
- 8) Marocco, A. and Pironneau, O., Optimum Design with Lagrangian Finite Elements — Design of an Electromagnet , Comp. Math. in Applied Mech. and Engineering, Vol. 15, 1978, pp. 217-308.
- 9) Delfour, M; Payre, G; Zolesio, J.P., Design of a Mass-Optimized Thermal Diffuser, in Proceedings of the NATO-ASI on Structural Optimization and Distributed Parameter Systems, Iowa City, 1980.
- 10) Cerepanov, G. P., Inverse Problems in the Plane Theory of Elasticity, FMM Vol. 38, pp. 963-979, 1974.
- 11) Banichuk, N.V. Optimality Conditions in the Problem of Seeking the Hole Shapes in Elastic Bodies", FMM Vol. 41, no. 1 pp. 920-925, 1977.
- 12) Wheeler, L, On the Role of Constant-stress Regions in the Problem of Minimizing Elastic Stress Concentration, J. Appl. of Solids and Structures, Vol. 12, pp. 779-789, 1976.
- 13) Leavitt, J. and Ungar, P., Circle Supports the Largest Sandpile, Comm. Pure and Appl. Math. Vol. 15, pp. 35-37, 1962.
- 14) Aronsson, G., An Integral Inequality and Plastic Torsion, Archive for Rational Mechanics and Analysis vol. 72, pp. 23-39, 1979.

- 15) Polya, G., Torsional Rigidity, Principal Frequency, Electrostatic Capacity, and Symmetrization, *Quarterly J. of Appl. Math.* vol. 6 pp. 267-277 (1948)
- 16) Murat, F., Contre-exemples pour divers problemes ou le controle intervient dans les coefficients, *Annali di Mat. Pura ed Appl.*, Ser. 4, vol. 112-~~113~~, 1977. **49-68**
- 17) Lurie, K.A., and Kłosowicz, B., "On the optimal nonhomogeneity of a torsional elastic bar", *Archive of Mechanics*, vol. 24, pp. 239-249, 1971.
- 18) Zavallani, A.; Maier, G.; Binda, L.; Shape Optimization of Plastic Structures by Zero-One Programming, *Proc. of IUTAM Symposium on Structural Optimization*, Warsaw; Springer-Verlag , 1975.
- 19) Cioranescu, D. and Saint Jean Paulin, J., Homogenization dans des Ouverts à cavités, *C.R. Acad. Paris* vol. 284, Ser. A. 857-860.
- 20) Spagnolo, S. Convergence in energy for elliptic equations, in *Numerical Solution of PDE's III*, Hubbard, ed., Academic Press, 1976.
- 21) Simon, L., On G-Convergence of Elliptic Operators, *Indiana Univ. Math. Journal*, vol. 28, pp. 567-594
- 22) Tartar, L. Estimation de coefficients homogenises, in *Springer Lecture Notes in Math.* vol. 704, pp. 364-373.
- 23) Helflin, Z. and Shtrikman, S., A Variational Approach to the theory of elastic behavior of multiphase materials, *J. Mech. and Phys. of Solids*, vol. 11, 1963, pp. 127-140.
- 24) Malpole, L.J. On bounds for the overall elastic moduli of inhomogeneous systems, I, II. *J. Mech. & Physics of Solids*, Vol. 14, pp. 151-162 and 289-301, 1966.
- 25) Hill, R., Elastic Properties of Reinforced Solids: Some Theoretical Principles, *J. Mech. and Physics of Solids*, Vol. 11, 1963, pp. 357-372.
- 26) Willis, J.R., Bounds and Self-Consistent Estimates for Overall Properties of Anisotropic Composites, *J. Mech. and Physics of Solids*, vol. 25, 1977, pp. 185-202.
- 27) Rozvany, G. I. N., *Optimal Design of Flexural Systems*, Pergamon Press New York, 1976.
- 28) Prager, W., *Introduction to Structural Optimization*; International Centre for Mechanical Sciences Courses and Lectures No. 212, Springer-Verlag, Vienna, 1974.

29) Temam, R. and Strang, G., Duality and Relaxation in the Variational Problems of Plasticity, *Journal de Mécanique*, to appear.

30) Cheng, K.-T., and Olhoff, N., An Investigation Concerning Optimal Design of Solid Elastic Plates, DCALM Report No. 174, The Danish Center for Applied Mathematics and Mechanics, March 1980.

31) Kohn, R., and Strang, G., Existence and Characterization of Optimal Geometries for some model problems in plastic limit analysis. To appear.

32) Strang, G., A Minimax Problem in Plasticity Theory, in Springer Lecture Notes in Mathematics Vol. 701, Springer-Verlag, 1979. 701

33) Ekeland, I., and Temam, R., Convex Analysis and Variational Problems, North-Holland, 1976.

34) Federer, H. Geometric Measure Theory, Springer-Verlag, 1969.

35) Kohn, R., and Strang, G., Optimal design for torsional rigidity, Proc. Symposium on Mixed and Hybrid Finite Element Methods, Atlanta, 1981.

36) Kohn, R., and Strang, G., Optimal design of cylinders in shear, MAFELAP Conference, Brunel, 1981.

37) Kohn, R., and Strang, G., Optimal design and convex analysis, in preparation (slowly).

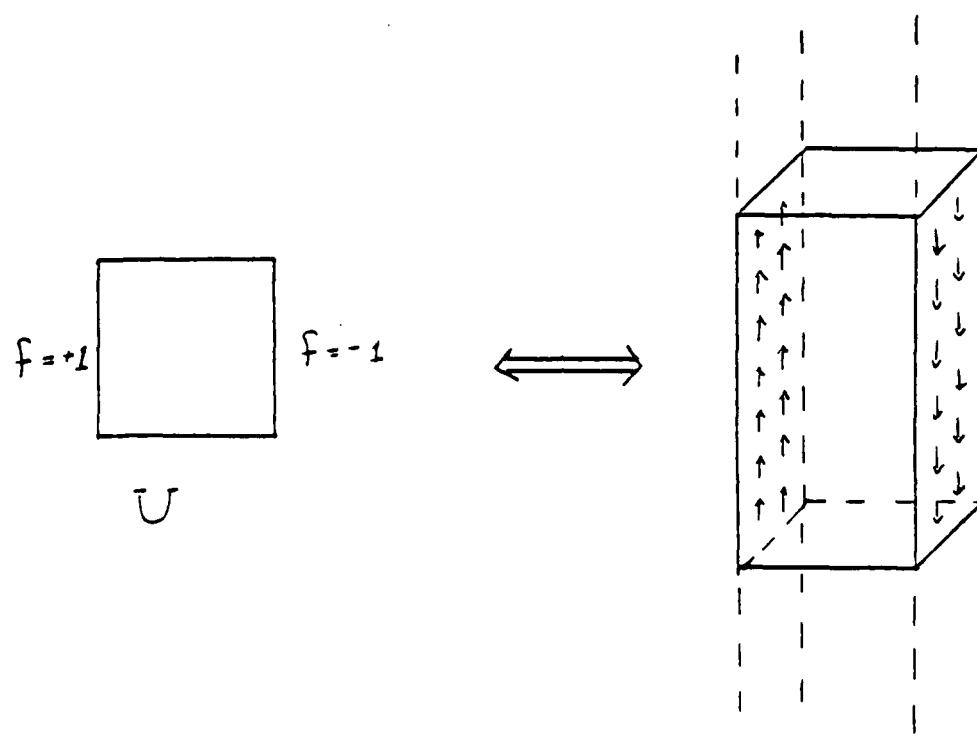


FIGURE 1

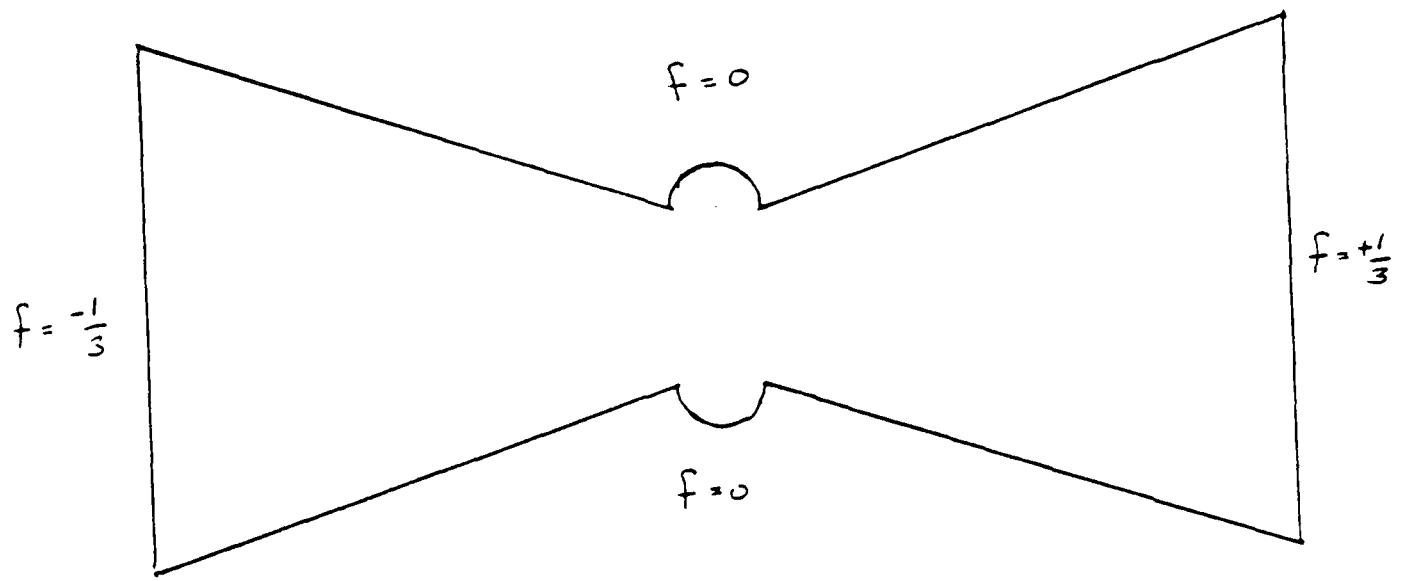
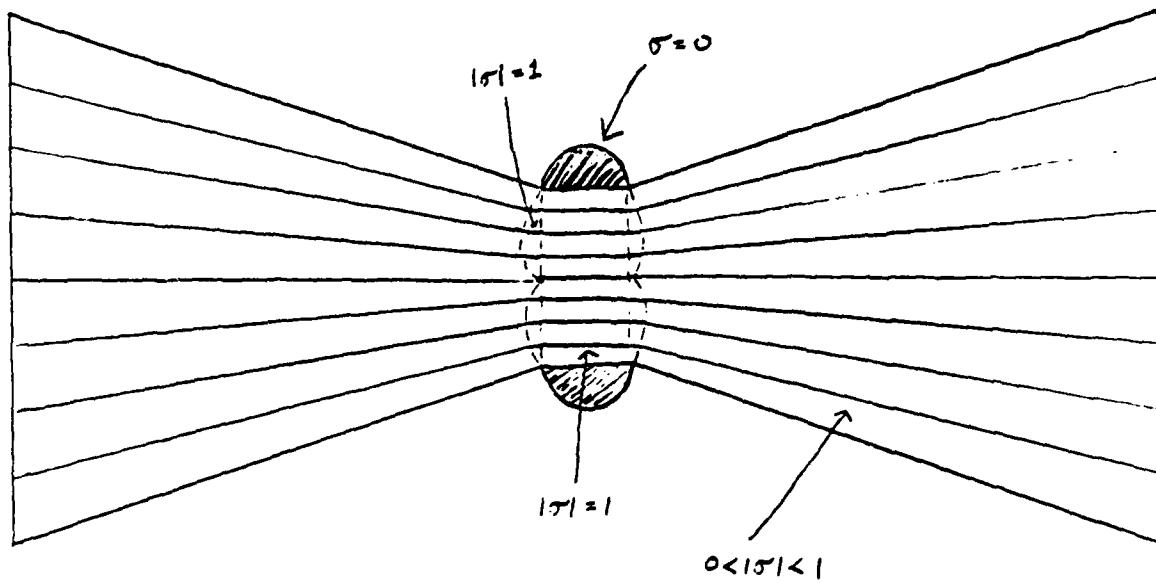
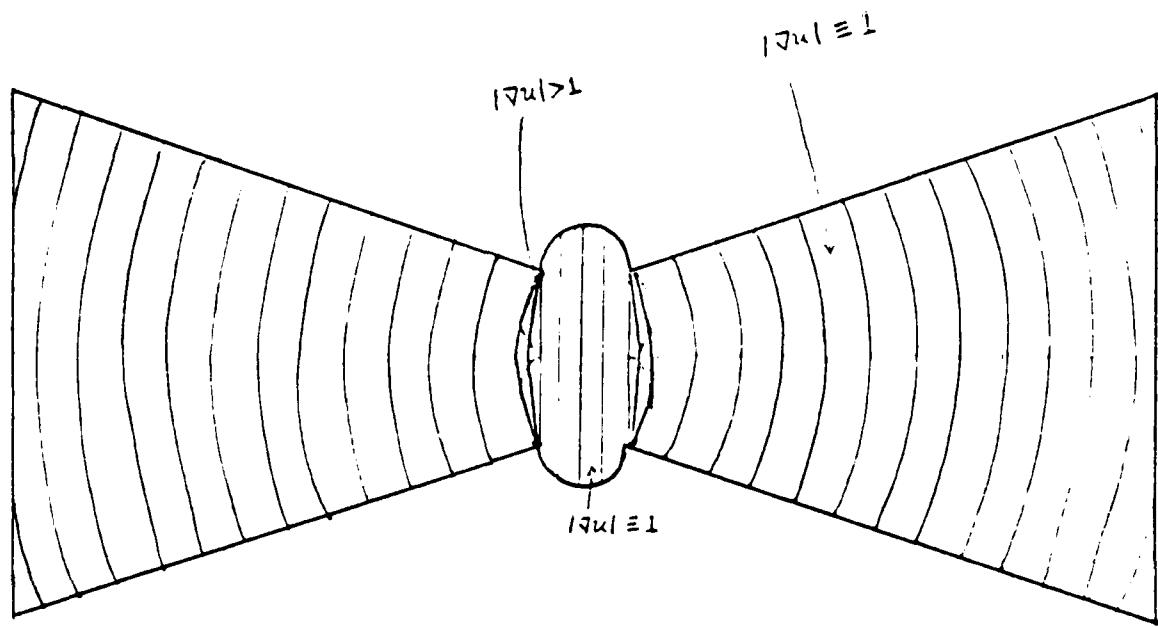
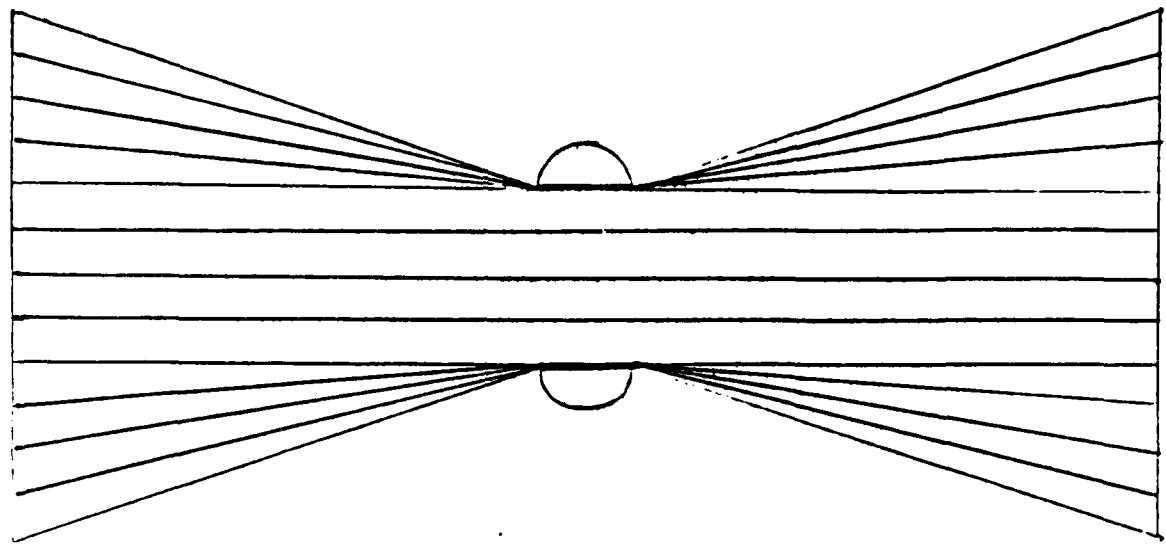


FIGURE 2

THE OPTIMAL STRESS FIELD  $\sigma$ FIGURE 3

THE DUAL DISPLACEMENT  $u$ FIGURE 4



LIMIT OF  $\tilde{E}_t^i \sigma_t$  AS  $t \rightarrow 0$

FIGURE 5

Stability and Error Bounds for a Fractional Step Scheme to Compute  
Weak Solutions to the Nonlinear Waterhammer Problem \*

Mitchell Luskin<sup>1</sup> and Blake Temple<sup>2</sup>  
Dept. of Mathematics Dept. of Mathematical Physics  
University of Michigan The Rockefeller University  
Ann Arbor, MI 48109 New York, NY 10021

ABSTRACT

A numerical method to compute weak solutions to an initial-boundary value problem for a nonlinear hyperbolic system which models fluid flow in a pipe is analyzed. The effect of friction is included by adding a quadratic zero order term to the system of conservation laws for compressible, frictionless flow. A priori bounds are obtained by means of a nonincreasing functional that is compatible with the friction effects and which is equivalent to the total variation of the solution. The boundary values for this problem cannot be imposed weakly, so new results on the regularity of the approximate solution at the boundary are given. Details will appear in the authors' paper "The existence of global weak solutions to the nonlinear waterhammer problem".

---

\* Presented at the University of Maryland on February 6, 1981,  
by Mitchell Luskin.

<sup>1</sup> Supported by the DOE under Contract DE-AC02-76ER03077 and by an NSF National Needs Postdoctoral Fellowship while the author was visiting at the Courant Institute of Mathematical Sciences.

<sup>2</sup> Supported by National Science Foundation Mathematical Sciences Postdoctoral Research Fellowship, Grant MCS80-17157.

## 1. Introduction

Fluid flow in pipelines is usually modeled by the quasilinear hyperbolic system

$$(1.1) \quad \begin{aligned} \rho_t + G_x &= 0, & (x,t) \in (0,1), \\ G_t + (G^2/\rho)_x + p(\rho)_x &= -f|G|G/2D\rho, \end{aligned}$$

where  $\rho$  is mass density,  $G$  is momentum density,  $p = p(\rho)$  is pressure,  $f = f(|G|)$  is the Moody friction factor, and  $D$  is pipe diameter. We shall give stability and error bounds for a numerical method to compute global weak solutions to (1.1) satisfying given initial conditions

$$(1.2) \quad \rho(x,0) = \rho_0(x), \quad G(x,0) = G_0(x), \quad x \in [0,1],$$

and given boundary conditions

$$(1.3a) \quad \rho(0,t) = \rho_B(t), \quad t \in (0,\infty),$$

$$(1.3b) \quad G(1,t) = 0.$$

This poses the classical "waterhammer" problem since the waterhammer phenomenon in hydraulics can be created by a sudden valve closure downstream (modeled by the boundary condition  $G(1,t) \equiv 0$ ) or by a rapid change in the pressure upstream (modeled by a discontinuity in  $\rho_B$ ). These events create pressure waves which are reflected at the boundaries.

The term  $-f|G|G/2D\rho$  accounts for the momentum loss due to viscous friction between the fluid and the pipe wall. Since the flow changes from laminar to turbulent at a flow rate near  $G_c = 2000\mu/D$  (where  $\mu$  is the dynamic viscosity), the properties of  $f$  also change at  $G = G_c$ . In the laminar regime

$$(1.4) \quad f(|G|) = 64\mu/|G|D, \quad |G| < G_c,$$

but the friction factor is determined experimentally for turbulent flow ( $|G| > G_c$ ) and depends on the pipe roughness (which we assume to be constant in space and time) as well as the flow rate. In particular, it can be observed from experimental data that there exists a constant  $f_1 > 0$  such that

$$(1.5) \quad \lim_{|G| \rightarrow \infty} f(|G|) = f_1.$$

Thus, the friction term  $f|G|G/2D\rho$  is nearly quadratic in  $G$  for turbulent flow. Our analysis assumes only the following properties for  $H(G) = f|G|G/2D$ :

$$(1.6) \quad H(0) = 0$$

$$(1.7) \quad H_G \geq \frac{H}{G} \geq 0$$

$$(1.8) \quad H \text{ is locally Lipschitz continuous.}$$

Property (1.6) states that there should be no friction when there is no flow. Property (1.7) states that the relative change in the friction (assuming that  $\rho$  is fixed) is

greater than the relative change in the flow rate. This is obviously valid in the laminar regime ( $H(G) = 64G/\mu D$ ) and in the completely turbulent regime ( $H(G) = f_0|G|G$ ). Our study of the Moody diagram [13, p. 297] has led us to assume its validity in general. Property (1.8) is justified for all flow rates,  $G$ , except possibly at the transition flow rate  $|G| = G_c$  (see [7] where  $f$  is allowed to be multi-valued at  $|G| = G_c$ ).

We also assume that the sound speed,  $c > 0$ , is constant, i.e.,

$$(1.9) \quad p'(\rho) = c^2.$$

This is valid for an ideal gas which is maintained at a constant temperature by heat exchange between the gas, the pipe wall, and the surrounding environment. For many physical problems property (1.9) is also a good approximation for modeling the flow of liquids.

In [7], Luskin has shown for the initial-value problem (1.1)-(1.2) that a unique, global smooth solution exists if the initial data are in an appropriate invariant region and if the first derivatives of the initial data are sufficiently small. However, if the first derivatives of the initial data are too large, then discontinuities can be shown to occur even when the data is smooth. (This can be done using a variant of Lax's ideas for the frictionless case [4]). To allow for more general data here, we need to

consider weak solutions of (1.1). We call  $\rho, G \in L^\infty(\Omega)$   
a weak solution of (1.1) if

$$(1.10) \quad \begin{aligned} & \iint_{\Omega} [\rho \phi_t + G \phi_x] dx dt = 0 \\ & \iint_{\Omega} [G \phi_t + (G^2/\rho + p(\rho)) \phi_x - (f|G|G/2D\rho)\phi] dx dt = 0 \end{aligned}$$

for all  $\phi \in C_0^\infty(\Omega)$ , where  $\Omega = (0,1) \times (0,\infty)$ .

We have proven the following existence theorem in [8]  
by using the properties of our approximate solution that  
will be given in this paper.

Theorem 1. Assume that properties (1.6)-(1.9) hold and  
that

$$(1.11) \quad \underset{t>0}{\text{Var}} \ln \rho_B + \underset{x \in [0,1]}{\text{Var}} \ln \rho_0 + \underset{x \in [0,1]}{\text{Var}} \frac{G_0}{c\rho_0} < \ln \frac{3+\sqrt{5}}{2} \approx 0.96.$$

Then there exists a weak solution  $\rho, G \in L^\infty(\Omega)$  to (1.1).

The initial values are satisfied in the sense that

$$(1.12) \quad \rho(\cdot, t), G(\cdot, t) \in \text{Lip}([0, \infty), L^1(0,1))$$

and

$$\lim_{t \rightarrow 0} \rho(\cdot, t) = \rho_0, \quad \lim_{t \rightarrow 0} G(\cdot, t) = G_0 \text{ in } L^1(0,1)$$

The boundary values are satisfied in the sense that for  
any  $T > 0$ ,

$$(1.13) \quad \rho(x, \cdot), G(x, \cdot) \in \text{Lip}([0,1], L^1(0,T)),$$

and

$$\lim_{x \rightarrow 0} \rho(x, \cdot) = \rho_B, \quad \lim_{x \rightarrow 1} G(x, \cdot) = 0 \text{ in } L^1(0,T).$$

(Here, e.g.,  $\rho(\cdot, t) \in \text{Lip}([0, \infty), L^1(0, 1))$  means that there exists a constant,  $C$ , such that

$$|\rho(\cdot, t_1) - \rho(\cdot, t_2)|_{L^1(0, 1)} \leq C|t_1 - t_2|$$

for all  $t_1, t_2 \in [0, \infty)$ .

Also, without loss of generality we assume that

$\rho_0(0) = \rho_B(0)$  and that  $G_0(1) = 0$  by redefining  $\rho_0(0)$  and  $G_0(1)$  if necessary. In this way the incompatibility of initial and boundary data is accounted for in the left hand side of (1.11) by allowing  $\lim_{x \rightarrow 0} \rho_0(x) \neq \rho_0(0)$  and  $\lim_{x \rightarrow 1} G_0(x) \neq 0$ .

The only purpose of (1.11) is to guarantee a priori that the flow remains subsonic, i.e.,

$$(1.14) \quad |v| < c \quad \text{for } (x, t) \in \Omega,$$

where  $v = G/\rho$  is the velocity of the flow. This, in turn, guarantees solvability when boundary conditions (1.3) are imposed.

In general, boundary value problems for (1.1) in which either the density or the flow rate is assigned at each boundary can be solved uniquely only when the characteristic speeds  $\lambda_1, \lambda_2$  satisfy  $\lambda_1 < 0, \lambda_2 > 0$ . Our problem is posed in Eulerian coordinates where the characteristic speeds are  $\lambda_1(u) = v - c, \lambda_2(u) = v + c$ ; so (1.14) is required for  $\lambda_1 < 0, \lambda_2 > 0$ . Earlier work on the construction of solutions to initial-boundary value problems has been done

by Nishida and Smoller [12] and Liu [5] for the "piston problem", but a priori bounds similar to (1.14) were not required there. This is because the piston problem is posed in Lagrangian coordinates where the boundaries move with the fluid. Thus for the piston problem,  $\lambda_1 = -c$ ,  $\lambda_2 = c$ , so  $\lambda_1 < 0$ ,  $\lambda_2 > 0$  is already guaranteed a priori.

Note also that boundary condition (1.3b) is a "natural" boundary condition and could have been imposed weakly by requiring that

$$\iint_{\Omega} [\rho \phi_t + G \phi_x] dx dt = 0$$

for all  $\phi \in C_0^\infty((0,1] \times (0,\infty))$ . However, the boundary condition (1.3a) is not a natural boundary condition, and it was necessary for us to give new results on the regularity of the solution at the boundary in [8] in order to make sense of boundary condition (1.3a). This problem, as well, did not arise in [5] or [12] since the boundary conditions for the piston problem are "natural" boundary conditions and can be imposed weakly.

Our method is a fractional step procedure. In the first part of each step we use Glimm's [1] method to approximate the solution of the system of conservation laws for frictionless flow. The second part of each step accounts for the effect of friction on the flow, and involves solving an O.D.E. that is determined by the zero order term. Liu [6] and Ying and Wang [15] have also given bounds for a frictional step method for some systems of conservation laws with zero

order terms. However, their analyses took account of only the magnitude and not the orientation of the vector field given by the zero order terms. These methods are inadequate for our purposes because the physical friction term  $f|G|G/2DP$  is quadratic in  $G$  at infinity; and solutions to systems of conservation laws with quadratic zero order terms will "blow up" in finite time if the associated vector field is allowed to have an arbitrary orientation. Moreover, the methods in [6] and [15] will not imply the a priori bound (1.14) unless the orientation of the vector field is considered. Thus, it is crucial that we found a nonlinear functional which is equivalent to the variation norm and which is nonincreasing on both of the fractional steps. Although the functional introduced by Nishida [10] is nonincreasing for the system of conservation laws, it is inadequate for our purposes since it can increase on the friction step. However, we have shown [8] that if the zero order term satisfies certain monotonicity conditions, then the functional given by Liu [5] is nonincreasing for both fractional steps. These monotonicity conditions are satisfied by the physical friction term in (1.1) when the flow is subsonic. Numerical results for the solution of (1.1) by this fractional step procedure have recently been reported by Marchesin and Paes-Lema [9].

More details and proofs for the results reported here can be found in our paper [8].

## 2. Solution of the Riemann Problem

The solution of the Riemann problem is the crucial element of our method. The Riemann problem is the initial value problem for data which is constant to the left and right of  $x = 0$ . We study the Riemann problem for the nonlinear hyperbolic system

$$(2.1) \quad u_t + F(u)_x = 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}^T,$$

where  $u = (\rho, G)^{tr}$ ,  $F(u) = (G, G^2/\rho + p(\rho))^{tr}$ , and  $p'(\rho) = c^2$ . The eigenvalues of  $dF$  are

$$(2.2) \quad \lambda_1(u) = v + c, \quad \lambda_2(u) = v + c,$$

with corresponding right eigenvectors

$$R_1(u) = (1, v-c)^{tr}, \quad R_2(u) = (1, v+c)^{tr}$$

The main existence result is that for initial data

$$(2.3) \quad u(x,0) = u_0(x) = \begin{cases} u_L & \text{if } x < 0, \\ u_R & \text{if } x > 0 \end{cases}$$

(we always assume  $\rho_L, \rho_R > 0$ ) there exists a unique solution  $u(x,t) = u(x/t)$  such that  $u(x/t)$  consists of constant states separated by "shock wave" and "rarefaction wave" solutions [10].

We first discuss the rarefaction wave solutions.

We note that a smooth solution  $u(\xi)$ ,  $\xi = x/t$ , must satisfy

$$[dF - \xi I] \dot{u}(\xi) = 0.$$

Hence, a smooth solution  $u(\xi)$  must satisfy  $\dot{u}(\xi) \in \text{span}\{R_\ell(u(\xi))\}$  and  $\xi = \lambda_\ell(u(\xi))$  for  $\ell = 1$  or  $\ell = 2$ . An  $\ell$ -rarefaction wave is a continuous solution,  $u(x/t)$ , whose values lie on an integral curve of the eigenvector  $R_\ell$ . The functions

$$s = \frac{v + c \ln \rho}{2}, \quad r = \frac{v - c \ln \rho}{2},$$

are Riemann invariants; i.e.,

$$(2.4) \quad \nabla_u s \cdot R_1 = 0, \quad \nabla_u r \cdot R_2 = 0.$$

Hence,  $s$  [resp.  $r$ ] is constant on an integral curve of  $R_1$  [resp.  $R_2$ ]. Thus, the  $\ell$ -rarefaction curves can be defined by

$$(2.5a) \quad R_1(u_L) = \{u_R \mid r(u_R) \geq r(u_L), \quad s(u_R) = s(u_L)\}, \\ = \{u_R \mid v_L - v_R = -cz \quad \text{for } z = \ln \rho_L - \ln \rho_R \geq 0\},$$

$$(2.5b) \quad R_2(u_L) = \{u_R \mid r(u_R) = r(u_L), \quad s(u_R) \geq s(u_L)\}, \\ = \{u_R \mid v_L - v_R = -cz \quad \text{for } z = \ln \rho_R - \ln \rho_L \geq 0\}.$$

A 1-shock wave [resp. 2-shock wave] of speed  $\sigma$  is a weak solution

$$(2.6) \quad u(x,t) = \begin{cases} u_L & \text{if } x/t \leq \sigma, \\ u_R & \text{if } x/t > \sigma \end{cases}$$

which satisfies the Lax entropy condition [3]

$$(2.7a) \quad \lambda_1(u_L) > \sigma > \lambda_1(u_R)$$

[resp.

$$(2.7b) \quad \lambda_2(u_L) > \sigma > \lambda_2(u_R) ] .$$

Since  $u$  is a weak solution, it must also satisfy the Rankine-Hugoniot jump condition

$$(2.8) \quad \sigma[u_L - u_R] = F(u_L) - F(u_R) .$$

By eliminating  $\sigma$  in (2.8) and applying the Lax entropy condition we obtain the following 1-shock wave curves

$$(2.9a) \quad S_1(u_L) = \{u_R \mid v_L - v_R = c(e^{-z/2} - e^{z/2}) \\ \text{for } z = \ln \rho_L - \ln \rho_R \leq 0\}$$

$$(2.9b) \quad S_2(u_L) = \{u_R \mid v_L - v_R = c(e^{-z/2} - e^{z/2}) \\ \text{for } z = \ln \rho_R - \ln \rho_L \leq 0\} .$$

Substituting in (2.8) gives

$$(2.10) \quad \sigma = v_L - c e^{-z/2} = v_R - c e^{z/2} , \quad z = \ln \rho_L - \ln \rho_R \leq 0$$

for a 1-shock and

$$(2.11) \quad \sigma = v_L + c e^{-z/2} = v_R + c e^{z/2} , \quad z = \ln \rho_R - \ln \rho_L \leq 0$$

for a 2-shock.

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

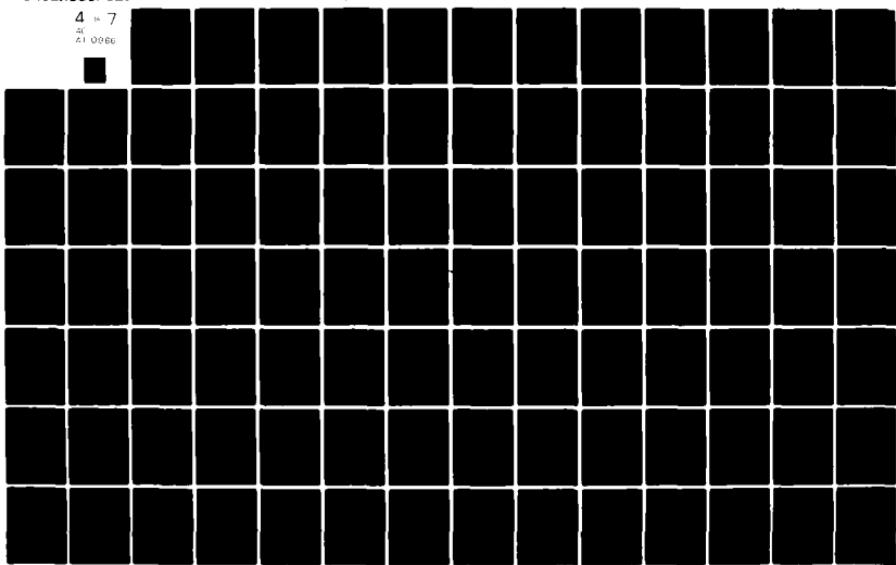
F/6 12/1

AFOSR-TR-82-0047

NL

UNCLASSIFIED

4 n 7  
41  
21 0066



It now easily follows from (2.8) and (2.9) that  $z = \ln \rho_L - \ln \rho_R$  [resp.  $z = \ln \rho_R - \rho_L$ ] is a regular parametrization of the  $C^2$  curve  $T_1(u_L) = R_1(u_L) \cup S_1(u_L)$  [resp.  $T_2(u_L) = R_2(u_L) \cup S_2(u_L)$ ]. We call  $z$  the "signed strength" of a given wave (so that the signed strength of a rarefaction wave is positive and the signed strength of a shock wave is negative), and we call  $|z|$  the strength of a wave. The existence theorem for Riemann problems follows directly from the fact that given any two states  $u_L$  and  $u_R$ , there exists a unique state  $u_M$  such that  $u_M \in T_1(u_L)$  and  $u_R \in T_2(u_M)$  [10]; i.e., the Riemann problem for (2.1) can always be uniquely solved by a 1-wave that connects  $u_L$  to  $u_M$  and a 2-wave that connects  $u_M$  to  $u_R$ .

Finally, we shall need to construct the solutions to certain initial-boundary value problems. When the boundary is  $x = 0$ , we consider the problem

$$(2.12) \quad \begin{aligned} u_t + F(u)_x &= 0, \quad (x,t) \in \mathbb{R}^T \times \mathbb{R}^T \\ u(x,0) &= u_0(x) = u_R, \quad x \in \mathbb{R}^T \\ \rho(0,t) &= \rho_L, \quad t \in \mathbb{R}^T \end{aligned}$$

It can be checked that there exists  $G_L$  such that  $u_R \in T_2(u_L)$ . But the 2-wave connecting  $u_L$  to  $u_R$  will take the value  $u_L$  at  $x = 0$  only if it has positive speed. If  $u_R \in R_2(u_L)$ , then it is necessary that  $\lambda_2(u) = v_L + c > 0$  to guarantee the 2-wave has positive speed. If

$u_R \in S_2(u_L)$ , then the corresponding 2-wave has positive speed if  $v_R > -c$ , since

$$\sigma > \lambda_2(u_R) = v_R + c > 0.$$

When the boundary is  $x = 1$ , we consider the problem

$$(2.13) \quad \begin{aligned} u_t + F(u)_x &= 0, & (x, t) \in \mathbb{R}^- \times \mathbb{R}^+, \\ u(x, 0) &= u_L, & x \in \mathbb{R}^- \\ G(0, t) &= 0, & t \in \mathbb{R}^+ \end{aligned}$$

In this case, there always exists  $\rho_R$  such that  $u_R = (\rho_R, 0)^T \in T_1(u_L)$ , but the 1-wave connecting  $u_L$  to  $u_R$  will take the value  $u_R$  at  $x = 1$  only if it has negative speed. This is true if  $v_L < c$ . Thus, the initial-boundary value problems (2.12) and (2.13) can be solved by simple waves if all the velocities occurring in the solution are subsonic.

### 3. Definition and Stability of the Fractional Step Scheme

In this section, we define our approximate solution to (1.1)-(1.3). Let  $h = 1/N$ ,  $N$  a positive integer,  $x_i = ih$ ,  $I_i = [x_{i-1}, x_i]$ ,  $I = [0, 1]$  and let  $k > 0$ ,  $t_j = jk$ ,  $J_j = [t_{j-1}, t_j]$ . Also, let  $\alpha = (\alpha_1, \dots)$ ,  $\alpha_i \in (0, 1)$ , be a sequence. We define approximate solutions  $\hat{u}_h = (\hat{\rho}_h, \hat{G}_h)^{tr}$  and  $u_h = (u_h, G_h)^{tr}$  inductively. Assume that  $\hat{u}_h$  and  $u_h$  are defined for  $t \leq t_j$ . Then  $\hat{u}_h$  on  $I \times J_{j+1}$  is the solution to

$$(3.1) \quad \begin{aligned} \hat{u}_{ht} + F(\hat{u}_h)_x &= 0 & (x, t) \in I \times J_{j+1} \\ \hat{\rho}_h(t, 0) &= \rho_B(t_{j+1}/2), & t \in J_{j+1}, \\ \hat{G}_h(t, 1) &= 0 & t \in J_{j+1}, \\ \hat{u}_h(x, t_j+) &= u_h(x_{i-1} + \alpha_j h, t_j^-), & x \in I_i, \text{ if } j > 0, \\ \hat{u}_h(x, 0+) &= u_0(x_{i-1}/2), & x \in I_i, \text{ if } j = 0 \end{aligned}$$

where  $F(u) = (G, G^2/\rho + p(\rho))^{tr}$  and  $u_0(x) = (\rho_0(x), G_0(x))^{tr}$ .

Next define the functions  $\check{u}(t, \check{u}) = (\check{\rho}(t, \check{u}), \check{G}(t, \check{u}))^{tr}$

by

$$(3.2) \quad \begin{aligned} \check{u}_t &= H(\check{u}), & t > 0 \\ \check{u}(0) &= \check{u} \end{aligned}$$

where  $H(u) = (0, -f|G|G/2D\rho)^{tr} \equiv (0, -H(G)/\rho)^{tr}$ .

Then we set

$$(3.3) \quad u_h(x, t) = \check{u}(t - t_j, \hat{u}_h(x, t)), \quad t \in J_{j+1}.$$

Note that (3.1) poses an initial value Riemann problem at each mesh point  $(x_i, t_j)$ ,  $0 < i < N$ , and a boundary Riemann problem of type (2.12) and (2.13) at  $i = 0$  and  $i = N$ , respectively. Therefore, we can use the Riemann problem solutions of Section 2 to solve (3.1) in  $J_{j+1}$ , so long as the waves in these solutions do not intersect in  $J_{j+1}$ , and so long as  $|v|$  remains less than  $c$  in the boundary problems. The results to be given in this section show that if (1.11) holds and if  $h/k \geq 4c$ , then  $|\hat{v}_h| < c$  in  $\Omega$  and the approximate solution can be successively constructed as described above on  $I \times J_j$  for  $j = 1, 2, \dots$ .

Note that if  $|\hat{v}_h| < c$  in  $I \times J_{j+1}$ , then  $|\lambda_i(\hat{u}_h)| < 2c$  in  $I \times J_{j+1}$ . In this case, the waves in the Riemann problem solutions do not interact since  $h/k \geq 4c$ . We shall always assume that  $h/k \geq 4c$  in the following.

We let  $\gamma_{ij}^l$  be the signed strength of the  $l$ -wave in the construction of  $\hat{u}_h$  which leaves  $(x_i, t_j)$ . Furthermore, define

$$L(j) = \sum_{i,l} |\gamma_{ij}^l|,$$

$$L_B(j) = \sum_{m>j} |\ln \rho_B(t_{m+1/2}) - \ln \rho_B(t_{m-1/2})|.$$

The following lemma shows that bounds on  $L(j) + L_B(j)$  imply bounds on the pointwise values of the solution.

Lemma 3.1. Suppose that

$$(3.4) \quad L_B(j) + L(j) \equiv v_j < \ln \frac{3+\sqrt{5}}{2}.$$

Then

$$(3.5) \quad \sup_{(x,t) \in I \times J_{j+1}} \left| \ln \frac{\hat{\rho}_h(x,t)}{\rho_\infty} \right| \leq v_j$$

and

$$\sup_{(x,t) \in I \times J_{j+1}} \left| \frac{\hat{v}_h(x,t)}{c} \right| \leq \frac{e^{v_j} - 1}{e^{v_j}/2} < 1$$

where  $\rho_\infty = \lim_{t \rightarrow +\infty} \rho_B(t)$ .

Lemma 3.1 implies that the waves in the Riemann problem solutions to (3.1) do not interact in  $I \times J_{j+1}$  if  $v_j < \ln (3+\sqrt{5})/2$  since  $k < h/4c$ . The next result is that  $v_j$  is nonincreasing for  $j = 1, 2, \dots$ . This implies that if  $v_0 < \ln (3+\sqrt{5})/2$ , then the approximate solutions can be constructed in  $I \times J_j$  for  $j = 1, 2, \dots$ .

Theorem 3.1. Suppose that (1.11) holds. Then

$$(3.7) \quad v_0 \leq v \equiv \var_{t>0} \ln \rho_B + \var_{x \in [0,1]} \ln \rho_0 + \var_{x \in [0,1]} \frac{v_0}{c} < \ln \frac{3+\sqrt{5}}{2}$$

and

$$(3.8) \quad v_{j+1} \equiv L_B(j+1) + L(j+1) \leq L_B(j) + L(j) \equiv v_j$$

for  $j = 0, 1, \dots$ .

#### 4. Regularity and Error Estimates

In this section, we first give regularity results for the approximate solution which show that  $u_h$  is  $L^1$  continuous in space and time to within an error dominated by the mesh length. These results are necessary to show that  $u_h$  converges to a solution  $u$  (after passing to a subsequence) that actually takes on the appropriate boundary values in the  $L^1$  sense.

Lemma 4.1. There exists a constant,  $C$ , such that

$$(4.1) \quad \int_0^1 |u_h(x, \tau_2) - u_h(x, \tau_1)| dx \leq C[|\tau_2 - \tau_1| + k].$$

Lemma 4.2. Assume that  $\alpha = (\alpha_1, \alpha_2, \dots)$  is equidistributed and  $T < \infty$ . Then there exists a constant,  $C = C(T)$ , such that for  $y_1, y_2 \in [0, 1]$ ,

$$(4.2) \quad \int_0^T |u_h(y_2, t) - u_h(y_1, t)| dt \leq C[|y_2 - y_1| + k]$$

for  $k$  sufficiently small.

To give an error estimate for our weak solution, we define

$$E(u_h, \phi) = - \int_0^T \int_0^1 [u_h \phi_t + F(u_h) \phi_x + H(u_h) \phi] dx dt$$

for  $\phi \in C_0^\infty((0, 1) \times (0, T))$ . Note that a weak solution,  $u$ , of (1.1) satisfies  $E(u, \phi) = 0$  for all  $\phi \in C_0^\infty((0, 1) \times (0, T))$ .

The following theorem gives a probabilistic measure of how much  $u_h$  varies from a weak solution.

Theorem 4.1. There exists a constant,  $C = C(T)$ , such that

$$(4.3) \quad \int E^2(u_h, \phi) d\alpha \leq Ch(|\phi|_\infty^2 + |\phi_x|_\infty^2)$$

The proofs of the estimates in this section and the argument giving the convergence of  $u_h$  to a weak solution of (1.1) can be found in our paper [8].

### Bibliography

- [1] Glimm, J., "Solutions in the large for nonlinear hyperbolic systems of equations," Comm. Pure Appl. Math., v. 18 (1965) 697-715.
- [2] Glimm, J. and P. D. Lax, "Decay of solutions of systems of nonlinear hyperbolic conservation laws," Mem. Amer. Math. Soc., v. 101, 1970.
- [3] Lax, P. D., "Hyperbolic systems of conservation laws, II," Comm. Pure Appl. Math., v. 10, 1957, pp. 537-566.
- [4] Lax, P. D., "Development of singularities of solutions of nonlinear hyperbolic partial differential equations," J. Math. Phys., v. 5, 1964, pp. 611-613.
- [5] Liu, T-P., "Initial-boundary value problems for gas dynamics," Arch. Rational Mech. Anal., v. 32, 1979, pp. 169-189.
- [6] Liu, T-P., "Quasilinear hyperbolic systems," Comm. Math. Phys., v. 68, 1979, pp. 141-172.
- [7] Luskin, M., "On the existence of global smooth solutions for a model equation for fluid flow in a pipe," J. Math. Anal. Appl., to appear.
- [8] Luskin, M. and B. Temple, "The existence of global weak solutions to the nonlinear waterhammer problem," preprint.
- [9] Marchesin, D. and P. J. Paes-Leme, "Shocks in gas pipelines," preprint.
- [10] Nishida, T., "Global solution for an initial boundary value problem of a quasilinear hyperbolic system,"

Proc. Jap. Acad., v. 44, 1968, pp. 642-646.

[11] Nishida, T. and J. Smoller, "Solutions in the large for some nonlinear hyperbolic conservation laws," Comm. Pure Appl. Math., v. 26, 1973, pp. 183-200.

[12] Nishida, T. and J. Smoller, "Mixed problems for nonlinear conservation laws," J. Diff. Eq., v. 23, 1977, pp. 244-269.

[13] Streeter, V. and E. B. Wylie, Fluid Mechanics, McGraw-Hill, New York, 1975, 6th ed.

[14] Temple, B., "Solutions in the large for some nonlinear hyperbolic conservation laws of gas dynamics," to appear in J. Diff. Eq.

[15] Ying, L-A., and C.-H. Wang, "Global solutions of the Cauchy problem for a nonhomogeneous quasilinear hyperbolic system," Comm. Pure Appl. Math., v. 33, 1980, pp. 579 - 597.

**Schauder Estimates for Finite Element Approximations  
on Second Order Elliptic Boundary Value Problems**

**Joachim A. Nitsche  
Institut für angewandte Mathematik  
Universität Freiburg, Deutschland**

### 0. Introduction

Let  $u$  be the solution of a second order elliptic boundary value problem and let  $u_h = R_h u \in S_h$  be the corresponding Ritz resp. finite element approximation onto the space  $S_h$ . Asking for  $L_\infty$ -estimates of  $u_h$  itself or the error  $u - u_h$  for approximation spaces  $S_h$  of order at least 3, i.e. for finite elements which are at least piecewise quadratics, the following results are to be mentioned:

(i) In Scott for  $N = 2$  dimensions it is proven

$$(0.1) \|u - R_h u\|_{L_\infty} \leq c h \inf_{\chi \in S_h} \|\nabla(u - \chi)\|_{L_\infty}$$

The proof is based on a carefull analysis of the approximability of the Green's function in the norm of  $W_1^1$ .

(ii) In Nitsche for arbitrary dimensions the a priori estimate

$$(0.2) \|R_h u\|_{L_\infty} + h \|\nabla(R_h u)\|_{L_\infty} \leq C \left\{ \|u\|_{L_\infty} + h \|\nabla u\|_{L_\infty} \right\}$$

was shown. Generalizing earlier results of Natterer the proof is based on the extensive use of certain weighted norms which are in the case of finite elements strongly connected with  $L_\infty$ -norms.

(iii) In Schatz - Wahlbin the estimate

$$(0.3) \|R_h u\|_{L_\infty} \leq C \|u\|_{L_\infty}$$

is proven. The method used is somehow between the other two

mentioned above.

The first aim of the present paper is to show that the estimate (0.3) can be derived directly following the lines of our former paper with the only difference that whenever the gradient of  $u$  enters the formulae then partial integration has to be applied. Actually this happens only in 3 places. In order to give a self-contained representation we repeat the arguments of our paper, the only changes are explained in Remark 5 and 6. For the sake of simplicity resp. clearness we give the analysis in the section 3 for the Laplacian serving as a model problem. The case of a general second order equation causes no additional difficulties , this is discussed in section 6. The proof of a crucial lemma was skipped in our former paper. It is given in detail in section 4.

The second aim of this paper is to show the boundedness of the Ritz operator in Hölder- resp. Lipschitz spaces. These spaces are the adequate ones in treating nonlinear elliptic problems. The boundedness of the Ritz operator in the corresponding norms at least simplifies the analysis of finite element procedures, in some cases it is essential.

Seemingly up to now Hölder spaces did not find any attention in the finite element literature. Corresponding to this a priori estimates or error estimates in the norms of these spaces do not exist in the literature.

## 1. Notations, Finite Elements

In the following  $\Omega \subseteq \mathbb{R}^N$  denotes a bounded domain with boundary  $\partial\Omega$  sufficiently smooth. For any  $\Omega' \subseteq \Omega$  let  $W_p^k(\Omega')$  be the Sobolev space of functions having  $L_p$ -integrable derivatives of order up to  $k$ . The norms are indicated by the corresponding subscripts. In the case  $p=2$  we also adopt  $H_k(\Omega') = W_2^k(\Omega')$ . The norms then are written shortly

$$(1.1) \quad \| \cdot \|_{k, \Omega'} = \| \cdot \|_{W_2^k(\Omega')} .$$

In addition we will use the abbreviation for boundary norms:

$$(1.2) \quad \| \cdot \|_{\partial\Omega'} = \| \cdot \|_{W_2^k(\partial\Omega')} .$$

Moreover  $\Omega'$  is skipped in case of  $\Omega' = \Omega$  and  $k$  in case of  $k=0$ .

The use of weighted norms resp. semi-norms will be essential. They are defined by

$$(1.3) \quad \|\nabla^k v\|_{\alpha, \Omega'} = \left\{ \sum_{|\gamma|=k} \iint_{\Omega'} \mu^{-\alpha} |\nabla^\gamma v|^2 dx \right\}^{1/2}$$

with  $\mu$  given by

$$(1.4) \quad \mu = \mu(x) = |x - x_0|^2 + \beta^2$$

$(x_0 \in \bar{\Omega}, q > 0)$ . The boundary semi-norms  $\| \cdot \|_{\partial \Omega}$ , are defined in the corresponding way.

By  $\Gamma_h$  a subdivision of  $\Omega$  into generalized simplices  $\Delta$  is meant, i. e.  $\Delta$  is a simplex if  $\Delta$  intersects  $\partial \Omega$  in at most a finite number of points and otherwise one of the faces may be curved.  $\Gamma_h$  is called  $K$ -regular if to any  $\Delta \in \Gamma_h$  there are two spheres of diameters  $K^{-1}h$  and  $h$  such that  $\Delta$  contains the one and is contained in the other.

The finite element spaces  $S_h = S(\Gamma_h)$  we will work with have the following structure: Let  $m$  being an integer fixed. Any element of  $S_h$  is continuous in  $\Omega$  and the restriction to  $\Delta \in \Gamma_h$  is a polynomial of degree less than  $m$ . In curved elements we use isoparametric modifications as discussed by CIARLET - RAVIART, ZLAMAL.  $S_h^o$  is the intersection of  $S_h$  and  $H_1^o$ , the closure in  $H_1$  of the functions with compact support.

By construction we have  $S_h \subseteq H_1$  but in general  $S_h \not\subseteq H_K$  for  $k \geq 2$ . It is useful to introduce the spaces  $H_K' = H_K'(\Gamma_h)$  consisting of functions the restriction of which to any  $\Delta$  is in  $H_K(\Delta)$ . Obviously  $S_h \subseteq H_K'$  for all  $k$ .

Parallel to above we use 'broken' seminorms

$$(1.5) \quad \|\nabla^k v\|' = \left\{ \sum_{\Delta \in \Gamma_h} \|\nabla^k v\|_{\Delta, \partial}^2 \right\}^{1/2}$$

$$\|\nabla^k v\|'_K = \left\{ \sum_{\Delta \in \Gamma_h} \|\nabla^k v\|_{\Delta, \partial}^2 \right\}^{1/2}.$$

## 2. Approximation Theory in weighted Norms

In the estimates of the next sections  $c, c_1, \dots$  etc. will denote generic constants which may differ at different locations. Unless otherwise stated they depend only on (i) the domain  $\Omega$ , (ii) the dimension  $N$ , (iii) the regularity parameter  $K$ , and (iv) the order  $m$ .

Essential is the fact that the function  $\mu$  (1.4) does not change too fast in any  $\Delta \in \mathcal{P}_h$  if  $\rho$  is not small compared with  $h$ :

Lemma 1: Let  $\rho \geq h$ . Then for any  $\Delta \in \mathcal{P}_h$

$$(2.1) \quad \bar{\mu}_\Delta = \sup_{x \in \Delta} \mu(x) \leq 6 \inf_{x \in \Delta} \mu(x) = \underline{\mu}_\Delta.$$

Proof: Let  $\bar{x}, \underline{x} \in \bar{\Delta}$  be points where  $\mu$  attains its maximum and minimum. Then

$$(2.2) \quad \bar{\mu}_\Delta = \mu(\bar{x}) = \mu(\underline{x}) + (\bar{x} - \underline{x}) \cdot \nabla \mu(\tilde{x}).$$

Now we have

$$(2.3) \quad |\nabla \mu(\tilde{x})| = 2 |\tilde{x} - x_0| \leq 2 \bar{\mu}_\Delta^{1/2}$$

and

$$(2.4) \quad |\bar{x} - \underline{x}| \leq h \leq \rho \leq \bar{\mu}_\Delta^{1/2}$$

leading to

$$(2.5) \quad \bar{\mu}_\alpha \leq \mu_\alpha + 2 \mu_\alpha^{1/2} \bar{\mu}_\alpha^{1/2}$$

$$\leq 3 \mu_\alpha + \frac{1}{2} \bar{\mu}_\alpha$$

#

Next let  $v \in C^0 \cap H_\rho'$  be given and  $\chi \in S_h$  an appropriate interpolation. Then the estimate

$$(2.6) \quad \|v^\kappa(v-\chi)\|_{L_2(\Omega)}^2 \leq c h^{2(\ell-k)} \|D^\ell v\|_{\ell(\Omega)}^2$$

for any  $\alpha \in P_h$  and  $0 \leq k < \ell \leq m$  is well known. Because of Lemma 1 we derive from this

$$(2.7) \quad \|v^\kappa(v-\chi)\|_{\alpha(\Omega)}^2 \leq c 6^{|\alpha|} h^{2(\ell-k)} \|D^\ell v\|_{\alpha(\Omega)}^2$$

The power  $\alpha$  will be within the range  $|\alpha| \leq N+1$ . Thus we drop the factor  $6^{|\alpha|}$ . Summation over all  $\alpha \in P_h$  gives

Lemma 2: Let  $\rho \geq h$ . To any  $v \in C^0 \cap H_\rho'$  there is a  $\chi \in S_h$  according to

$$(2.8) \quad \|v^\kappa(v-\chi)\|_{\alpha}^2 \leq c h^{\ell-k} \|D^\ell v\|_{\alpha}^2$$

for  $0 \leq k < 1 \leq n$ .

Remark 1: Since (2.9) is valid also for  $v \in C^0 \cap H_\rho' \cap H_1^0$  with  $\chi \in S_h^0$  the lemma remains valid in this situation.

For any  $w \in H_1(\alpha)$  the trace theorem gives

$$(2.9) \quad \|w\|_{L_2(\partial\Delta)}^2 \leq c \left\{ h^{-1} \|w\|_{L_2(\alpha)}^2 + h \|\nabla w\|_{L_2(\alpha)}^2 \right\}.$$

Using the arguments of above we get

Corollary 2: Under the assumptions of Lemma 2

$$(2.10) \quad |\nabla^\ell (\gamma - \chi)|' \leq c h^{\ell-k-\frac{1}{2}} \|\nabla^\ell v\|_\alpha'$$

is valid in addition.

The proof of the next lemma and corollary follows the same lines and is omitted here.

Lemma 3: For  $\chi \in S_h$  and  $0 \leq k < l \leq n$  inverse relations of the type

$$(2.11) \quad \|\nabla^\ell \chi\|_\alpha' \leq c h^{-(\ell-k)} |\nabla^k \chi|'_\infty$$

hold true.

Corollary 3: In addition to (2.11)

$$(2.12) \quad |\nabla^\ell \chi|'_\alpha \leq c h^{-(\ell-k+\frac{1}{2})} \|\nabla^k \chi\|_\alpha'$$

holds true. Here  $k = 1$  is accepted.

In the subsequent sections we will apply these approximation results to functions  $V$  of the structure  $V = f^{-\alpha} \varphi$  with  $\varphi \in S_h$ . Then a certain super-approximability property holds:

Lemma 4: Let  $\varphi \in S_h$  be given. The function  $f^{-\alpha} \varphi$  can be approximated by an element  $\chi \in S_h$  according to

$$(2.13) \quad \begin{aligned} h \|\nabla^2(f^{-\alpha}\varphi - \chi)\|_{-\alpha}^{''} + \|\nabla(f^{-\alpha}\varphi - \chi)\|_{-\alpha} & \\ + h^{1/2} \|\nabla(f^{-\alpha}\varphi - \chi)\|_{-\alpha}^{'} & \leq C \frac{h}{g} (\|\varphi\|_{\alpha+1} + \|\nabla\varphi\|_{\alpha}). \end{aligned}$$

Proof: We apply Lemma 2 and Corollary 2 with  $l = m$  and get the bound

$$(2.14) \quad C h^{m-1} \|\nabla^m(f^{-\alpha}\varphi)\|_{-\alpha}^{'}$$

for the three terms on the left hand side in (2.13). Since  $\varphi$  is piecewise a polynomial of degree less than  $m$  and because of

$$(2.15) \quad |Df^{-\alpha}| \leq C f^{-\alpha - 1/2}$$

Leibniz' rule gives

$$(2.16) \quad \|\nabla^m(f^{-\alpha}\varphi)\|_{-\alpha}^{' \prime} \leq C \sum_{n=0}^{m-1} \|\nabla^n \varphi\|_{\alpha+m-n}^{'}$$

Now we apply Lemma 3 for the terms with  $n \geq 1$ :

$$(2.17) \quad \begin{aligned} \|\nabla^m (\tilde{\mu}^{-\alpha} \varphi)\|_{-\alpha}' &\leq \\ &\leq C \left\{ \|\varphi\|_{\alpha+m} + \sum_1^{m-1} h^{1-n} \|\nabla \varphi\|_{\alpha+m-n} \right\}. \end{aligned}$$

Using finally the obvious inequality for  $\beta > 0$

$$(2.18) \quad \| \cdot \|_{(\cdot)+\beta} \leq \| \cdot \|_{(\cdot)} \varsigma^{-\beta}$$

we end up with

$$(2.19) \quad \begin{aligned} \|\nabla^m (\tilde{\mu}^{-\alpha} \varphi)\|_{-\alpha}' &\leq \\ &\leq C \left\{ \varsigma^{1-m} \|\varphi\|_{\alpha+1} + \sum_1^{m-1} h^{1-n} \varsigma^{n-m} \|\nabla \varphi\|_{\alpha} \right\} \end{aligned}$$

and therefore

$$(2.20) \quad \begin{aligned} C h^{m-1} \|\nabla^m (\tilde{\mu}^{-\alpha} \varphi)\|_{-\alpha}' &\leq \\ &\leq C \left\{ (h/\varsigma)^{m-1} + \sum_1^{m-1} (h/\varsigma)^{n-m} \right\} \left\{ \|\varphi\|_{\alpha+1} + \|\nabla \varphi\|_{\alpha} \right\}. \end{aligned}$$

The first brackets on the right hand side are bounded by  $m h/\varsigma$   
since  $h \leq \varsigma$  is assumed.

As was pointed out in the introduction weighted norms are strongly  
connected with the  $L_{\infty}$ -norm. First we show

Lemma 5: Let  $\alpha > \frac{N}{2}$ . Then for any  $v \in L_\infty$  it is

$$(2.21) \quad \|v\|_\alpha^2 \leq C \int_{-\infty}^{+\infty} \|v\|_\infty^2 r^{-2\alpha+N} dr.$$

Proof: We can estimate

$$(2.22) \quad \|v\|_\alpha^2 \leq \|v\|_\infty^2 \int_Q r^{-\alpha} dx$$

and further with  $r$  denoting the distance  $|x - x_0|$

$$(2.23) \quad \begin{aligned} \int_Q r^{-\alpha} dx &\leq C \int_0^\infty (r^2 + q^2)^{-\alpha} r^{N-1} dr \\ &\leq C \int_0^\infty (r+q)^{N-1-2\alpha} dr. \end{aligned}$$

For elements in the space  $S_h$  there is the counterpart;

Lemma 6: Let  $\alpha > \frac{N}{2}$  and  $h \leq q$ . Then for  $x \in S_h$  the inequality

$$(2.24) \quad \|x\|_{L_\infty}^2 \leq c \int_{x_0 \in Q} h^{-N} \sup_{x \in Q} \|x\|_\alpha^2$$

holds true.

Proof: Let  $r_0 \in \bar{Q}$  be chosen such that

$$(2.25) \quad x(x_0) = \pm \|x\|_{L_\infty}$$

and let  $\Delta_0$  be (one of) the simplices with  $x_0 \in \bar{\Delta}_0$ .

$\chi$  restricted to  $\Delta_0$  is a polynomial of finite degree, i. e. an element of a finite dimensional space. In this case any two norms are equivalent. Since  $\Delta_0$  is of size  $h$  there is a constant  $c$  depending only on  $K$ ,  $N$ , and  $m$  such that

$$(2.26) \quad \|\chi\|_{L_\infty(\Delta_0)}^2 \leq c h^{-N} \|\chi\|_{L_2(\Delta_0)}^2.$$

Because of the choice of  $x_0$  it is

$$(2.27) \quad g^2 \leq \mu(x) \leq g^2 + h^2 \leq 2g^2.$$

Therefore we get further

$$(2.28) \quad \begin{aligned} \|\chi\|_{L_\infty(\Delta_0)}^2 &\leq c h^{-N} g^{2\alpha} \|\chi\|_{K, \Delta_0}^2 \\ &\leq c h^{-N} g^{2\alpha} \|\chi\|_K^2. \end{aligned}$$

Remark 2: The last two lemmata show that the  $\alpha$ -norm and the  $L_\infty$ -norm are equivalent in the spaces  $S_h$ .

### 3. The Boundedness of the Ritz Projection

In this section we restrict ourselves to the model problem

$$(3.1) \quad \begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

The weak formulation is:

$$(3.2) \quad \begin{aligned} \text{Find } u \in H_1^0, \quad \text{such that} \\ D(u, v) &= (f, v) \end{aligned}$$

holds for all  $v \in H_1^0$ .

Here  $D(\cdot, \cdot)$  denotes the Dirichlet integral

$$(3.3) \quad D(u, v) = (\nabla u, \nabla v) = \iint \sum u_{,i} v_{,i} \, dx.$$

The Ritz-approximation  $\varphi = \varphi_h \in S_h^0$  is characterized by the relation

$$(3.4) \quad D(\varphi, x) = (f, x) \quad \text{for } x \in S_\Omega^0$$

or alternately by

$$(3.5) \quad D(\varphi, x) = D(u, x) \quad \text{for } x \in S_\Omega^0.$$

Remark 3: Throughout this section the letter  $\varphi$  denotes the Ritz-approximation on  $u$ .

In the first step of our analysis we derive a bound for the gradient of  $\varphi$  in a weighted norm. It is

$$\begin{aligned} \|\nabla \varphi\|_{\alpha} &= (\nabla \varphi, \mu^{-\alpha} \nabla p) \\ (3.6) \quad &= D(p, \mu^{-\alpha} p) - (\nabla \varphi, p \nabla \mu^{-\alpha}) \\ &= D(p, \mu^{-\alpha} p) + \frac{1}{2} \int \varphi^2 \Delta \mu^{-\alpha}. \end{aligned}$$

Because of

$$(3.7) \quad \Delta \mu^{-\alpha} \leq C \mu^{-\alpha-1}$$

we get

$$(3.8) \quad \|\nabla \varphi\|_{\alpha}^2 \leq D(p, \mu^{-\alpha} p) + C \|\varphi\|_{\alpha+1}^2.$$

Next we use the identity

$$\begin{aligned} (3.9) \quad D(p, \mu^{-\alpha} p) &= D(p, \mu^{-\alpha} p - x) - \\ &\quad - D(x, \mu^{-\alpha} p - x) + D(x, \mu^{-\alpha} p) \end{aligned}$$

valid for any  $x \in S_Q$  because of (3.5). By the aid of Schwarz' inequality in the form

$$(3.10) \quad |D(v, w)| \leq \|\nabla v\|_{\alpha} \|\nabla w\|_{-\alpha}$$

and Lemma 4 we find for the first term on the right hand side of

$$(3.11) \quad \begin{aligned} |D(u^{-\kappa}\varphi - \chi, \varphi)| &\leq c \frac{\hbar}{\varrho} \left\{ \| \nabla \varphi \|_{\kappa} + \| \varphi \|_{\alpha+1} \right\} \| \nabla \varphi \|_{\kappa} \\ &\leq c \frac{\hbar}{\varrho} \left\{ \| D\varphi \|_{\kappa}^2 + \| \varphi \|_{\alpha+1}^2 \right\}. \end{aligned}$$

Our aim is to avoid any derivatives of  $u$  in the estimates.

Therefore we have to apply partial integration in order to handle the two other terms in (3.9). We get

$$(3.12) \quad \begin{aligned} D(u, \tilde{\mu}^{-\kappa}\varphi - \chi) &= \sum_{\alpha \in \mathbb{N}_0^n} \int_{\Omega} u (\tilde{\mu}^{-\kappa}\varphi - \chi) d\alpha \\ &\quad - \sum_{\alpha \in \mathbb{N}_0^n} \int_{\Omega} u \alpha (\tilde{\mu}^{-\kappa}\varphi - \chi) dx \end{aligned}$$

which may be estimated by

$$(3.13) \quad \begin{aligned} |D(u, \tilde{\mu}^{-\kappa}\varphi - \chi)| &\leq \| u \|_{\kappa}' \| \tilde{\mu}^{-\kappa}\varphi - \chi \|_{-\kappa}' \\ &\quad + \| u \|_{\kappa} \| \alpha (\tilde{\mu}^{-\kappa}\varphi - \chi) \|_{-\alpha}'. \end{aligned}$$

If  $\chi$  is chosen according to Lemma 4 then

$$(3.14) \quad \begin{aligned} |D(u, \tilde{\mu}^{-\kappa}\varphi - \chi)| &\leq c \frac{\hbar}{\varrho} \left\{ \| \nabla \varphi \|_{\kappa} + \| \varphi \|_{\alpha+1} \right\} \\ &\quad \left\{ \hbar^{-1/2} \| u \|_{\kappa}' + \hbar^{-1} \| u \|_{\kappa} \right\}. \end{aligned}$$

In order to shorten the formulae we introduce

$$(3.15) \quad N_{\kappa}(u) := \left\{ \hbar^{-2} \| u \|_{\kappa}^2 + \hbar^{-2} \| u \|_{\kappa}'^2 + \| u \|_{\alpha+1}^2 \right\}^{1/2}.$$

Then we come to - note  $\delta \leq q$  -

$$(3.16) \quad |D(u, \mu^{-\alpha} \varphi - \chi)| \leq c \frac{\delta}{q} \left\{ \| \nabla \varphi \|_{\alpha}^2 + \| \varphi \|_{\alpha+1}^2 \right\} \\ + c N_{\alpha}(u)^2.$$

Following the same line but this time using Lemma 3 and Corollary 3 we get

$$(3.17) \quad |D(u, \mu^{-\alpha} \varphi)| \leq c \left\{ \| \nabla \varphi \|_{\alpha}^2 + \| \varphi \|_{\alpha+1}^2 \right\} N_{\alpha}(u).$$

Schwarz' inequality in the form

$$(3.18) \quad |AB| \leq \delta A^2 + \frac{1}{4\delta} B^2$$

for  $0 < \delta < 1$  leads to

$$(3.19) \quad |D(u, \mu^{-\alpha} \varphi)| \leq \delta \left\{ \| \nabla \varphi \|_{\alpha}^2 + \| \varphi \|_{\alpha+1}^2 \right\} \\ + \frac{c}{\delta} N_{\alpha}(u)^2.$$

Now we combine (3.9), (3.11), (3.16) and (3.19) with (3.8).

This gives

$$(3.20) \quad \| \nabla \varphi \|_{\alpha}^2 \leq \left( C_1 \frac{\delta}{q} + \delta \right) \| \nabla \varphi \|_{\alpha}^2 \\ + c \| \varphi \|_{\alpha+1}^2 + \frac{c}{\delta} N_{\alpha}(u)^2.$$

We choose  $\delta = 1/3$  and impose the condition on  $\varphi$

$$(3.21) \quad \varphi \geq \gamma_1 h \quad \text{with } \gamma_1 = \max(1, 3c_1).$$

Then we get

$$(3.22) \quad \|\nabla \varphi\|_{\alpha}^2 \leq c_2 \|\varphi\|_{\alpha+1}^2 + c N_{\alpha}(h)^2.$$

Remark 4: In (3.20) we used for the constant in front of  $\|\nabla \varphi\|_{\alpha}$  the numbering  $c_1$ , since this special constant appeared in the condition (3.21). Similarly the constant  $c_2$  in front of  $\|\varphi\|_{\alpha+1}$  appears in a further condition.

Remark 5: In the analysis given in we did not use partial integration. There  $\|\nabla u\|_{\alpha}$  enters instead of  $N_{\alpha}(h)$ .

In the second step we introduce the auxiliary function  $w$  defined by

$$(3.23) \quad -\Delta w = \int_{\Omega}^{-\alpha-1} \varphi \quad \text{in } \Omega, \\ w = 0 \quad \text{on } \partial\Omega.$$

The reason is obvious since then

$$(3.24) \quad \|\varphi\|_{\alpha+1}^2 = D(\varphi, w)$$

which may be rewritten with  $x \in S_{\epsilon}^0$  arbitrary

$$(3.25) \quad \|\varphi\|_{\alpha+1}^2 = D(\varphi, w-x) - D(u, w-x) + D(u, w).$$

Using the definition of  $w$  we get at once for the last term on the right hand side

$$(3.26) \quad \begin{aligned} D(u, w) &= (u, \varphi)_{\alpha+1} \\ &\leq \delta \|\varphi\|_{\alpha+1}^2 + \frac{1}{4\delta} \|u\|_{\alpha+1}^2. \end{aligned}$$

Using (3.22) we get for the first term with  $0 < \delta < 1$

$$(3.27) \quad \begin{aligned} |D(\varphi, w-x)| &\leq \|D\varphi\|_{\alpha}^2 \|D(w-x)\|_{-\alpha} \\ &\leq \delta \|\varphi\|_{\alpha}^2 + \frac{1}{4\delta} \|D(w-x)\|_{-\alpha}^2 \\ &\leq C_2 \delta \|\varphi\|_{\alpha+1}^2 + C \|V(u)\|_{\alpha}^2 + \frac{1}{4\delta} \|D(w-x)\|_{-\alpha}^2. \end{aligned}$$

Finally the middle term on the right hand side of (3.25) has to be treated by partial integration. Similar to above we come to

$$(3.28) \quad \begin{aligned} |D(u, w-x)| &\leq \|u\|_{\alpha}^1 \|D(w-x)\|_{-\alpha}^1 + \\ &\quad + \|u\|_{\alpha}^1 \|D^2(w-x)\|_{-\alpha}^1 \\ &\leq N_{\alpha}(u)^2 + h \|V(w-x)\|_{-\alpha}^1 + h^2 \|V^2(w-x)\|_{-\alpha}^1. \end{aligned}$$

By means of the last three estimates we derive from (3.25)

$$\begin{aligned}
 \| \varphi \|_{\alpha+1}^2 &\leq (1 + c_2) \delta \| \varphi \|_{\alpha+1}^2 + \frac{c}{\delta} N_\alpha(u)^2 \\
 (3.29) \quad &+ \frac{c}{\delta} \left\{ h \|\nabla(w-x)\|_{-\alpha}^2 + h \|\nabla(w-x)\|_{-\alpha}^2 \right. \\
 &\quad \left. + h^2 \|\nabla^2(w-x)\|_{-\alpha}^2 \right\}.
 \end{aligned}$$

The choice  $\delta = 2 + 2c_2$  leads to

$$\begin{aligned}
 \| \varphi \|_{\alpha+1}^2 &\leq c N_\alpha(u)^2 + c \left\{ h \|\nabla(w-x)\|_{-\alpha}^2 \right. \\
 (3.30) \quad &\quad \left. + h \|\nabla(w-x)\|_{-\alpha}^2 + h^2 \|\nabla^2(w-x)\|_{-\alpha}^2 \right\}.
 \end{aligned}$$

Remark 6: The counterpart of the last inequality in our former analysis was

$$(3.31) \quad \| \varphi \|_{\alpha+1}^2 \leq c \| u \|_{\alpha+1}^2 + c \| P u \|_{\alpha}^2 + c \| \nabla(w-x) \|_{-\alpha}^2.$$

The third step consists in analyzing the terms with  $w-x$  in (3.30) which still depend on  $\varphi$  since  $w$  does. Since  $\varphi$  and hence  $\mu^{-\alpha-1} \varphi$  is in  $H_1$ , the shift theorem guarantees  $w \in H_3$ . We have assumed  $n \geq 3$ , i.e. at least quadratic finite elements are used. Therefore we may choose according to Lemma 2 and Corollary 2 with  $l = 3$  and get from (3.30)

$$(3.32) \quad \| \varphi \|_{\alpha+1}^2 \leq c N_\alpha(u)^2 + c h^2 \| \nabla^3 w \|_{-\alpha}^2.$$

The next section is devoted to the proof of

Lemma 7: Let  $\alpha$  be in the range  $N/2 < \alpha < (N+1)/2$ . Then for any  $v \in H_1 \cap H_3$  with  $\Delta w \in H_1^\alpha$  the a priori estimate

$$(3.33) \quad \|\nabla^3 w\|_{-\alpha} \leq \|\nabla \Delta w\|_{-\alpha} + c \beta^{-2} \|\Delta w\|_{-\alpha-1}$$

holds true.

Because of the definition of  $w$  (3.23) we find

$$(3.34) \quad \|\Delta w\|_{-\alpha-1} = \|\varphi\|_{\alpha+1}$$

and

$$(3.35) \quad \begin{aligned} \|\nabla \Delta w\|_{-\alpha} &= \|\nabla (\beta^{-\alpha-1} \varphi)\|_{-\alpha} \\ &\leq c \left\{ \|\varphi\|_{\alpha+3} + \|\nabla \varphi\|_{\alpha+2} \right\} \\ &\leq c \beta^{-2} \left\{ \|\varphi\|_{\alpha+1} + \|\nabla \varphi\|_{\alpha} \right\}. \end{aligned}$$

Now using (3.22) we derive from (3.32)

$$(3.36) \quad \|\varphi\|_{\alpha+1} \leq C_3 \frac{h^2}{\beta^2} \|\varphi\|_{\alpha+1} + c N_\alpha(u).$$

In analogy to (3.21) we impose the side constraint

$$(3.37) \quad \beta \geq \gamma_2 h \quad \text{with } \gamma_2 = \max(\gamma_1, \sqrt{2} C_3)$$

on  $\varphi$ . This leads to

Theorem 8: For  $\alpha \in (N/2, (N+1)/2)$  and under the condition

$\varphi \geq \gamma_2 h$  the  $(\alpha+1)$ -norm of the Ritz-approximation  $\varphi = R_h u$   
is bounded by the composed  $\alpha$ -norm  $N_\alpha(\cdot)$  of  $u$  itself

$$(3.38) \quad \|\varphi\|_{\alpha+1} \leq C N_\alpha(u)$$

with  $C$  independent of  $h$ ,  $\varphi$  and the point  $x_0$ .

## 4. Proof of Lemma 7

The general shift theorem in the theory of elliptic equations includes the two statements

Let  $v \in \overset{\circ}{H}_1 \cap H_2$ . Then

$$(4.1) \quad \|\nabla^2 v\| \leq C \|\Delta v\|.$$

Let  $v \in \overset{\circ}{H}_1 \cap H_3$ . Then

$$(4.2) \quad \|\nabla^3 v\| \leq C \{ \|\nabla \Delta v\| + \|\Delta v\| \}.$$

A direct consequence is

Lemma 9: Let  $v \in \overset{\circ}{H}_1 \cap H_2$  resp.  $v \in \overset{\circ}{H}_1 \cap H_3$ . The in weighted norms for  $\beta$  arbitrary

$$(4.3) \quad \|\nabla^2 v\|_{\beta} \leq C \{ \| \Delta v \|_{\beta} + \|\nabla v\|_{\beta+1} + \| v \|_{\beta+2} \},$$

$$(4.4) \quad \|\nabla^3 v\|_{\beta} \leq C \{ \|\nabla \Delta v\|_{\beta} + \|\Delta v\|_{\beta+1} + \\ + \|\nabla v\|_{\beta+2} + \| v \|_{\beta+3} \}$$

are valid.

Proof: We will give the details only for (4.3), the second case is handled in the same way. For convenience we use  $\varepsilon = \beta/2$ . We can rewrite the integrand in

$$(4.5) \quad \|\nabla^2 v\|_{\beta}^2 = \sum_{i,k} \iint (\mu^{-\epsilon} v_{ik})^2 dx$$

by

$$(4.6) \quad \begin{aligned} \mu^{-\epsilon} v_{ik} &= (\mu^{-\epsilon} v)_{ik} - v_i (\mu^{-\epsilon})_k - \\ &\quad - v_k (\mu^{-\epsilon})_i - v (\mu^{-\epsilon})_{ik}. \end{aligned}$$

Therefore we get using (2.15)

$$(4.7) \quad \|\nabla^2 v\|_{\beta} \leq 2 \|\nabla(\mu^{-\epsilon} v)\| + C (\|\nabla v\|_{\beta+1} + \|v\|_{\beta+2}).$$

In the similar way it is

$$(4.8) \quad \Delta(\mu^{-\epsilon} v) = \mu^{-\epsilon} \Delta v + 2 \nabla v \cdot \nabla(\mu^{-\epsilon}) + v \Delta \mu^{-\epsilon}$$

leading to

$$(4.9) \quad \|\Delta(\mu^{-\epsilon} v)\| \leq 2 \|\Delta v\|_{\beta} + C (\|\nabla v\|_{\beta+1} + \|v\|_{\beta+2}).$$

(4.7) together with (4.9) gives (4.3).

After these preparations we go back to the function  $w$  defined

by (3.23) and the a priori estimate stated in Lemma 7. By

Lemma 9 we have

$$(4.10) \quad \begin{aligned} \|\nabla^3 w\|_{-\alpha} &\leq C \left\{ \|\nabla \Delta w\|_{-\alpha} + \|\Delta w\|_{-\alpha+1} + \right. \\ &\quad \left. + \|\nabla w\|_{-\alpha+2} + \|w\|_{-\alpha+3} \right\}. \end{aligned}$$

We have at once

$$(4.11) \quad \|\Delta W\|_{-\alpha+1} \leq \beta^{-2} \|DW\|_{-\alpha-1} .$$

In order to complete the proof of Lemma 7 we have to show that the sum

$$(4.13) \quad \|\nabla W\|_{-\alpha+2} + \|W\|_{-\alpha+3}$$

is bounded by the right hand side of (3.33). Our choice of  $\alpha$  leads to

$$(4.14) \quad \frac{1}{2}(3-N) < -\alpha+2 < \frac{1}{2}(4-N) .$$

Therefore the weight  $-\alpha+2$  of the term  $\nabla W$  in (4.13) is positive in case of  $N = 2, 3$  dimensions and negative for  $N \geq 4$  dimensions. Moreover in case of  $N = 3$  dimensions we have

$$(4.15) \quad 0 < -\alpha+2 < \frac{N}{2} - 1 .$$

According to this the cases of 2, 3 or higher dimension have to be treated separately. This will be clearer because of the following

Lemma 10: Let  $v \in H_1 \cap H_2$ . Then

(i) for  $\beta < 0$  the norms  $\|\nabla v\|_{\beta}$  and  $\|v\|_{\beta+1}$  are comparable modulo  $\|\Delta v\|_{\beta-1}$ , i.e.  $\approx$ .

$$(4.16) \quad \begin{aligned} \|\nabla v\|_{\beta} &\leq C \left\{ \|v\|_{\beta+1} + \|\Delta v\|_{\beta-1} \right\}, \\ \|v\|_{\beta+1} &\leq C \left\{ \|\nabla v\|_{\beta} + \|\Delta v\|_{\beta-1} \right\}, \end{aligned}$$

(ii) for  $0 < \beta < \frac{N}{2} - 1$  ( $N > 2$ ) both terms are bounded by the last, i.e.  $\approx$ .

$$(4.17) \quad \|\nabla v\|_{\beta} + \|v\|_{\beta+1} \leq C \|\Delta v\|_{\beta-1}.$$

Proof: The identity

$$(4.18) \quad \|\nabla v\|_{\beta}^2 = (v, \mu^{-\beta} v) - \iint_{\Omega} v \nabla v \cdot \nabla \mu^{-\beta} dx$$

leads to

$$(4.19) \quad \|\nabla v\|_{\beta}^2 = (v, -\Delta v)_{\beta} + \frac{1}{2} \iint_{\Omega} v^2 \Delta \mu^{-\beta} dx.$$

direct differentiation gives  $-t = |x - x_0|$

$$(4.20) \quad \Delta \mu^{-\beta} = -2\beta \mu^{-\beta-2} \{ N\beta^2 + (N-2\beta-2)r^2 \}.$$

Thus in case (i)  $\Delta \mu^{-\beta}$  is bounded from above and below by  $C \mu^{-\beta-1}$  giving

$$(4.21) \quad \begin{aligned} \| \nabla v \|_{\beta}^2 &\leq (v, -\Delta v)_{\beta} + \bar{C} \| v \|_{\beta+1}^2, \\ &\geq (v, -\Delta v)_{\beta} + C \| v \|_{\beta+1}^2 \end{aligned}$$

This proves (4.16) since

$$(4.22) \quad |(v, -\Delta v)_{\beta}| \leq \delta \| v \|_{\beta+1}^2 + \frac{1}{4\delta} \| \Delta v \|_{\beta-1}^2.$$

In case (ii) we have

$$(4.23) \quad \Delta \mu^{-\beta} \leq -C' \mu^{-\beta-1}$$

with a positive constant  $c'$  giving

$$(4.24) \quad \begin{aligned} \| \nabla v \|_{\beta}^2 + c' \| v \|_{\beta+1}^2 &\leq (v, -\Delta v)_{\beta} \\ &\leq \frac{1}{2} c' \| v \|_{\beta+1}^2 + \frac{1}{2c'} \| \Delta v \|_{\beta-1}^2. \end{aligned}$$

We are now able to give a short proof of Lemma 7 for  $N = 3$  dimensions. Because of (4.15) and the second part of Lemma 10 we have

$$(4.25) \quad \begin{aligned} \| \nabla w \|_{-\alpha+2} + \| w \|_{-\alpha+3} &\leq C \| \Delta w \|_{-\alpha+1} \\ &\leq C g^{-2} \| \Delta w \|_{-\alpha-1}. \end{aligned}$$

Now let us consider the case of  $N = 2$  dimensions. We will give an explicit proof of

Corollary 2: Under the assumptions of Lemma 2 the terms  $\|\nabla v\|_{\beta+2}$  in (4.3) resp.  $\|\nabla v\|_{\beta+3}$  in (4.4) can be dropped in case of  $N = 2$  dimensions, provided  $\Delta v \in H_1^0$ .

Remark 7: The restriction to  $N = 2$  dimensions is unnecessary.

But we will need it only in this case.

Before we give the proof let us finish the proof of Lemma 7.

We need now - see (4.13) - a bound of  $\|\nabla w\|_{-\alpha+2}$  only. In the present case we have  $1/2 < -\alpha+2 < 1$ . Let  $p_2 > 2$  be fixed. Then  $\alpha-1 < 2/p_2$ . Now we apply Hölder's inequality with  $p = p_2/2 > 1$  and get

$$(4.26) \quad \begin{aligned} \|\nabla w\|_{-\alpha+2}^2 &= \iint \mu^{\alpha-2} |\nabla w|^2 dx \\ &\leq \|\nabla w\|_{L_{p_2}}^2 \left\{ \iint \mu^{(\alpha-2)q} dx \right\}^{1/q} \end{aligned}$$

with  $1/q = 1 - 1/p$ . Direct calculation - see the proof of Lemma 5 - leads to

$$(4.27) \quad \|\nabla w\|_{-\alpha+2} \leq c \rho^{-\lambda} \|\nabla w\|_{L_{p_2}}$$

with

$$(4.28) \quad \lambda = 1 - \alpha + 2/p_2$$

Next let  $p_1$  be defined by

$$(4.29) \quad 1/p_1 = 1/2 + 1/p_2$$

By the aid of standard a priori estimates - see Morrey pp. 80 and 157 - we get

$$(4.30) \quad \|\nabla w\|_{L^{p_2}} \leq C \|\nabla^2 w\|_{L^{p_1}}$$

and

$$(4.31) \quad \|\nabla^2 w\|_{L^{p_1}} \leq C \|\Delta w\|_{L^{p_1}}.$$

In our case we have  $1 < p_1 < 2$ . Therefore we may apply once more Hölder's inequality to

$$(4.32) \quad \|\Delta w\|_{L^{p_1}}^{p_1} = \left\| \left( \mu^{\alpha+1} |\Delta w|^2 \right)^{p_1/2} \right\|_{L^{\mu}}^{\mu - (\alpha+1)p_1/2}$$

this time with  $p = 2/p_1$ . Similar to above we get

$$(4.33) \quad \|\Delta w\|_{L^{p_1}} \leq C \int_0^\mu \|\Delta w\|_{-\alpha-1}^{\mu} d\mu$$

with

$$(4.35) \quad \mu = 1 + \alpha - \frac{2}{p_2}.$$

The combination of (4.27), (4.30), (4.31), and (4.33) leads to

$$(4.35) \quad \|\nabla w\|_{-\alpha+2} \leq C \delta^{-2} \|\Delta w\|_{-\alpha-1}$$

what finishes the proof of Lemma 7 for  $N = 2$  dimensions.

We will later on need the trace theorem in weighted norms in the form

Lemma 11: Let  $v \in H_1$ . Then for  $\delta > 0$

$$(4.36) \quad |v|_{\beta+1/2} \leq \delta \|\nabla v\|_\beta + C(\delta^{-1}) \|v\|_{\beta+1}.$$

Proof: (4.36) is shown by applying the standard trace theorem

$$(4.37) \quad |V|^2 \leq c \left\{ \|V\|^2 + \|V\| \|\nabla V\| \right\}$$

to  $V = \mu^{-\beta/2 - 1/4} v$ .

Proof of Corollary 9: In  $N = 2$  dimensions - we denote the variables by  $x, y$  - it is

$$(4.38) \quad \begin{aligned} |\nabla^2 v|^2 - |\Delta v|^2 &= -2 \left( v_{xx} v_{yy} - v_{xy}^2 \right) \\ &= -2 \left\{ (v_y v_{xx})_y - (v_y v_{xy})_x \right\} \end{aligned}$$

and therefore

$$(4.39) \quad \begin{aligned} \|\nabla^2 v\|_\rho^2 - \|\Delta v\|_\rho^2 &= 2 \int_Q t^{-\rho} v_y d\nu_x \\ &\quad + 2 \iint_Q v_y \left\{ v_{xx} (t^{-\rho})_y - v_{xy} (t^{-\rho})_x \right\} dx dy \end{aligned}$$

resp.

$$(4.40) \quad \begin{aligned} \| \nabla^2 v \|_p^2 - \| \Delta v \|_p^2 &\leq 2 \oint_{\partial Q} \mu^{-p} v_y dV_x + \\ &+ C \| \nabla v \|_{p+1} \| \nabla^2 v \|_p. \end{aligned}$$

In order to analyze the boundary integral we introduce the arc length  $s$  and the angle  $\gamma = \gamma(s)$  between the tangent and the  $x$ -axis. Further  $v_s, v_n$  denote the tangential and normal differentiation. Because of  $v = 0$  on  $\partial Q$  we have

$$(4.41) \quad v_x = - \sin \gamma v_n, \quad v_y = \cos \gamma v_n$$

and with  $K = \gamma'$  being the curvature of  $\partial Q$

$$(4.42) \quad 2v_y dV_x = -2 \left\{ K \cos^2 \gamma v_n^2 + \sin \gamma \cos \gamma v_n v_{ns} \right\} ds.$$

We insert this in the boundary integral and apply partial integration because of  $v_{n ns} = (v_n)_s^2/2$ . Then we get

$$(4.43) \quad \left| \oint_{\partial Q} \mu^{-p} v_y dV_x \right| \leq C \| \nabla v \|_{p+1/2}^2.$$

With the help of Lemma 11 then (4.40) leads to

$$(4.44) \quad \| \nabla^2 v \|_p^2 - \| \Delta v \|_p^2 \leq 2 \delta \| \nabla^2 v \|_p^2 + \frac{C}{\delta} \| \nabla v \|_{p+1}^2.$$

This proves (4.3) without the last term on the right hand side.

In proving the second part of Corollary 9 we will skip some of the details. In the corresponding way to above we get the counterpart of (4.40)

$$(4.45) \quad \left\| \nabla^3 v \right\|_{\beta}^2 - \left\| \nabla \Delta v \right\|_{\beta}^2 \leq 2 \oint_{\Omega} \mu^{-\rho} (v_{yy} - v_{xx}) dV_{xy} + C \left\| \nabla^2 v \right\|_{\beta+1} \left\| \nabla^3 v \right\|_{\beta} .$$

On  $\partial \Omega$  we have for  $v$  arbitrary with the abbreviations

$$s := \sin \varphi, \quad c := \cos \varphi$$

$$(4.46) \quad \begin{aligned} v_{ss} &= c^2 v_{xx} + 2sc v_{xy} + s^2 v_{yy} + K v_n, \\ v_{ns} &= -sc v_{xx} + (c^2 - s^2) v_{xy} + sc v_{yy} - K v_s, \\ v_{nn} &= s^2 v_{xx} - 2sc v_{xy} + c^2 v_{yy}. \end{aligned}$$

The condition  $v = 0$  implies  $v_s = v_{ss} = 0$ . In addition

$\Delta v = 0$  implies  $v_{nn} = -v_n$ . Therefore we derive

$$(4.47) \quad \begin{aligned} v_{yy} - v_{xx} &= -2K \cos 2\varphi v_n + 2 \sin 2\varphi v_{ns} \\ 2v_{xy} &= 2K \sin 2\varphi v_n + 2 \cos 2\varphi v_{ns} \end{aligned}$$

Similar to above we then get

$$(4.48) \quad \left| \oint_{\Omega} (\mu^{-\rho} (v_{yy} - v_{xx})) dV_{xy} \right| \leq C \left\{ \left\| \nabla^2 v \right\|_{\beta+1/2}^2 + \left\| \nabla v \right\|_{\beta+1/2}^2 \right\}$$

and therefore with Lemma 11

$$(4.49) \quad \begin{aligned} \|\nabla^3 v\|_{\beta}^2 - \|\nabla \Delta v\|_{\beta}^2 &\leq 2\delta \|\nabla^3 v\|_{\beta}^2 + \\ &+ \frac{\epsilon}{\delta} \left\{ \|\nabla^2 v\|_{\beta+1}^2 + \|\nabla v\|_{\beta+2}^2 \right\} \end{aligned}$$

resp.

$$(4.50) \quad \|\nabla^3 v\|_{\beta} \leq C \left\{ \|\nabla \Delta v\|_{\beta} + \|\nabla^2 v\|_{\beta+1} + \|\nabla v\|_{\beta+2} \right\}.$$

Now we have to apply the first part of Corollary 9 to the second term on the right hand side of (4.50). #

Remark 8: Above we derived the a priori estimates needed for functions sufficiently smooth only. For instance (4.38) holds only for functions having third derivatives. By compactness arguments the validity of the estimates for functions with the stated regularity is shown.

The case of  $N \geq 4$  dimensions hardly is of practical importance. Therefore we give only an outline of the proof for these cases. In view of Lemma 10 and because of (4.44) it is only necessary to bound  $\|w\|_{-\alpha+3}$  in terms of  $\|\Delta w\|_{-\alpha-1}$  i. e. to find an upper bound of

$$(4.51) \quad \lambda(\alpha) = \sup_{-\alpha+3} \|w\|^2 / \|\Delta w\|_{-\alpha-1}^2$$

where the supremum is to be taken over all  $v \in H_1^0 \cap H_2^0$ .

Obviously the supremum is attained for an eigenfunction of the problem

$$(4.52) \quad \begin{aligned} \Delta (\mu^{\alpha+1} \Delta v) &= \lambda \mu^{\alpha-3} v \text{ in } \Omega, \\ v = \Delta v &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

In this way we ask for a lower bound of the smallest eigenvalue of problem (4.52). By standard arguments the monotonicity of  $\lambda$  with respect to the domain, i.e.  $\lambda(\Omega_1) \leq \lambda(\Omega_2)$  in case of  $\Omega_1 \subseteq \Omega_2$ , is shown. Therefore an upper bound for  $\lambda(\Omega)$  is given by the corresponding  $\lambda$  for the ball with center in  $x_0$  and radius  $d = \text{diameter } (\Omega)$ . The eigenfunktion corresponding to the lowest eigenvalue then depends only on  $r = |x - x_0|$  (or at least one does).

Using the representation

$$(4.53) \quad \nabla v \sim v' = r^{1-N} \int_0^r s^{N-1} \Delta v ds$$

we get without difficulties

$$(4.54) \quad \|\nabla v\|_{-\alpha+2} \leq C \int_0^r s^{-2} \|\Delta v\|_{-\alpha-1} ds$$

which in view of Lemma 10 bounds  $\|\nabla v\|_{-\alpha+3}$  in the same way.

### 5. The Boundedness of the Ritz Approximation in Hölder Spaces

The Laplacian like any elliptic operator is not one to one with respect to the spaces  $C^k = C^k(\Omega)$  consisting of functions having continuous derivatives up to order  $k$  in  $\Omega$ . We will also abbreviate  $C = C^0$  and denote by  $\overset{\circ}{C}$  the space of continuous functions vanishing on the boundary  $\partial\Omega$ . Of course the image  $f = -\Delta u$  of any  $u \in \overset{\circ}{C} \cap C^{k+2}$  ( $k > 0$ ) is in  $C^k$  but to  $f \in C^k$  there may not be an original  $u \in \overset{\circ}{C} \cap C^{k+2}$  as is demonstrated in two dimensions by the counterexample

$$(5.1) \quad u = (x^2 - y^2) / \ln(x^2 + y^2)^{1/2}$$

with  $\Omega$  the unit sphere.

The situation is changed in case of Hölder- (resp. Lipschitz-) spaces. These spaces, denoted by  $C^{k,\lambda} = C^{k,\lambda}(\Omega)$  with  $\lambda$  according to  $0 < \lambda \leq 1$ , consist of all functions  $k$ -times continuously differentiable such that the highest derivatives are Hölder-continuous to the exponent  $\lambda$ . In  $C^{k,\lambda}$  a norm is given by

$$(5.2) \quad \|v\|_{C^{k,\lambda}} = \sum_{|\beta| \leq k} \|D^\beta v\|_{L_\infty} + \sum_{|\beta|=k} [D^\beta v]_\lambda$$

with

$$(5.3) \quad [w]_\lambda = \sup_{\substack{x,y \in Q \\ x \neq y}} \frac{|u(x) - u(y)|}{|x-y|^\lambda}.$$

Equipped with this norm  $C^{k,\lambda}$  is a Banach space. The Laplacian is a one to one mapping of  $\overset{\circ}{C} \cap C^{k+2,\lambda}$  into  $C^{k,\lambda}$ . Especially

$$(5.4) \quad \|u\|_{C^{k+2,\lambda}} = \|(-\Delta)^{1/2} f\|_{C^{k,\lambda}} \leq c \|f\|_{C^{k,\lambda}}.$$

Such an a priori estimate is referred to 'Schauder estimate'.

The aim of this section is the proof of corresponding estimates with  $u$  replaced by  $\varphi = R_h u$  being the Ritz approximation.

A first result in this direction is more or less a direct consequence of Theorem 8. See the proof of Lemma 5 - the right hand side of (3.38) is bounded by

$$(5.5) \quad N_\alpha(u) \leq c \varrho^{-\alpha+N/2} h^{-1} \|u\|_{L_\infty}.$$

By Lemma 6 we know

$$(5.7) \quad \|\varphi\|_{L_\infty} \leq c \varrho^{\alpha+1} h^{-N/2} \sup_{x_0} \|\varphi\|_{\alpha+\eta}.$$

Besides (3.37)  $\varrho$  is arbitrary. Now we fix  $\varrho = \gamma_2 h$  and get

$$(5.8) \quad \| \varphi \|_{L_\infty} = \| R_h u \|_{L_\infty} \leq C \| u \|_{L_\infty} .$$

This gives

Theorem 12: The Ritz operator is bounded as mapping of  $C$  into itself.

The spaces  $C^{k,\lambda}$  are compactly embedded on  $C$ . There is a general principle to bound the norm in  $C^{k,\lambda}$  of a linear projection operator by means of the norm in  $C$  which we will discuss now. The situation is that we have two Banach spaces  $X_1, X_2$  (with norms  $\| \cdot \|_1, \| \cdot \|_2$ ) with  $X_2$  compactly embedded in  $X_1$ . Further we have a collection  $\{S_h \mid 0 < h \leq 1\}$  of subspaces of  $X_2$ . Let approximation- and inverse-quantities

$\sigma_h$  and  $\tau_h$  be introduced according to

(A) To any  $y \in X_2$  there is a  $\gamma \in S_h$  such that simultaneously

$$(5.9) \quad \begin{aligned} \| y - \gamma \|_1 &\leq \sigma_h \| y \|_2 , \\ \| y \|_2 &\leq c_1 \| y \|_1 \end{aligned}$$

is valid with  $c_1$  independent of  $h$ .

(I) For any  $x \in S_h$  a Bernstein type inequality holds

$$(5.10) \quad \| x \|_2 \leq \tau_h \| x \|_1 .$$

We will 'say' the collection  $\{S_h\}$  fulfills the AI-condition if

$$(5.11) \quad K := \sup_h \rho_h \tau_h < \infty.$$

Remark: Under 'reasonable' assumptions  $\rho_h$  will tend to zero with  $h$ . For finite dimensional spaces  $S_h$  the quantities  $\tau_h$  are finite since then any two norms are equivalent. With  $h \rightarrow 0$  resp.  $\dim(S_h) \rightarrow \infty$  then  $\tau_h$  will also tend to infinity. The AI-condition just balances this.

The mentioned principle is

Lemma 13: Let  $X_1, X_2$  be as described above and  $\{S_h\}$  a collection of subspaces of  $X_2$ . Further let  $\{P_h : X_1 \rightarrow S_h\}$  be a collection of linear projection operators of  $X_1$  onto  $S_h$  which are uniformly bounded as mappings of  $X_1$  into itself.

i. e.

$$(5.12) \quad \|P_h\|_1 = \sup_{y \neq 0} \frac{\|P_h y\|_1}{\|y\|_1} \leq p_1$$

with  $p_1$  independent of  $h$ . If  $\{S_h\}$  fulfills the AI-condition then  $\{P_h\}$  as mapping of  $X_2$  into itself is uniformly bounded with

$$(5.13) \quad \|P_h\|_2 = \sup_{y \neq 0} \frac{\|P_h y\|_2}{\|y\|_2} \leq p_2 := (c_1 + 3K)p_1.$$

Proof: Because of  $X_2 \subseteq X_1$  and  $S_h \subseteq X_2$  of course  $P_h$  is a linear projection of  $X_2$  into itself. Let  $y \in X_2$  be given and  $\tilde{y} \in S_h$  be chosen according to (5.9). Then

$$(5.14) \|P_h y\|_2 \leq \|P_h y - \tilde{y}\|_2 + c_1 \|y\|_2 .$$

Since  $P_h y - \tilde{y}$  is an element of  $S_h$  we may apply (5.10) getting

$$(5.15) \begin{aligned} \|P_h y\|_2 &\leq \tau_h \|P_h y - \tilde{y}\|_1 + c_1 \|y\|_2 \\ &\leq \tau_h \{ \|P_h y - y\|_1 + \|y - \tilde{y}\|_1 \} + c_1 \|y\|_2 . \end{aligned}$$

Now we use the inequality

$$(5.16) \|y - P_h y\|_1 \leq (1 + \|P_h\|_1) \inf_{\tilde{y} \in S_h} \|y - \tilde{y}\|_1$$

the proof of which - in order to give a selfcontained presentation - is as follows: Let  $\tilde{y} \in S_h$  be arbitrary. Because of  $P_h \tilde{y} = \tilde{y}$  we have

$$(5.17) \begin{aligned} \|y - P_h y\|_1 &= \|(y - \tilde{y} - P_h(y - \tilde{y}))\|_1 \\ &\leq (1 + \|P_h\|_1) \|y - \tilde{y}\|_1 . \end{aligned}$$

In (5.16) resp. (5.17) we may use on the right hand side  $\tilde{y} = \tilde{y}$

Because of the assumption (5.12) we get from (5.15)

$$(5.18) \quad \|y - P_h y\|_1 \leq (1 + p_1) \|y - g\|_1$$

and

$$(5.19) \quad \|P_h y\|_2 \leq (2 + p_1) \tau_h \|y - g\|_1 + c_1 \|y\|_2.$$

Finally using (5.9) we come to

$$(5.20) \quad \|P_h y\|_2 \leq \{(2 + p_1) \tau_h + c_1\} \|y\|_2.$$

The norm of any projection operator is bounded from below by 1. Therefore we can also bound

$$(5.21) \quad p_2 \leq (3k + c_1) p_1$$

which is more convenient.

Remark 10: Lemma 13 first was stated in Nitsche

It remains to prove

Lemma 14: Assume  $S_h \subseteq C^k$ . Then with  $X_1 = C$  and  $X_2 = C^{k+1}$  the finite element spaces  $S_h^\circ$  fulfill the AI-condition.

The consequence is the final result:

Theorem 15: Assume  $S_h \subseteq C^k$ . Then the Ritz operator is bounded as mapping of  $C^{k+1}$  into itself.

Proof of Lemma 14: The finite elements discussed in section 2 are only in  $C$ . We will give the proof only for the case  $k = 0$ . The case  $k \geq 1$  follows the same lines and is omitted here in order to avoid the introduction of finite elements with higher smoothness. We will show that the standard interpolation will have the properties needed. Especially we will show

$$(5.22) \quad \sigma_h \leq c h^\lambda, \quad \tau_h \leq c h^{-\lambda}.$$

First we prove the estimate for  $\tau_h$ . Similar to Lemma 3 we have for  $x \in S_h$

$$(5.23) \quad \|\nabla x\|'_{L_\infty} = \max_{\alpha \in P_h} \|\nabla x\|_{L_\infty(\alpha)} \leq c h^{-1} \|x\|_{L_\infty}.$$

Now let  $x, y \in Q$  be given. In case of  $|x-y| \geq h$  we have trivially

$$(5.24) \quad \frac{|x(x)-x(y)|}{|x-y|^\lambda} \leq 2 h^{-\lambda} \|x\|_{L_\infty}.$$

In case of  $|x-y| < h$  we come from

$$(5.25) \quad |x(x)-x(y)| \leq |x-y| \|\nabla x\|'_{L_\infty}$$

to

$$(5.26) \quad \frac{|x(z) - x(y)|}{|z-y|^{\lambda}} \leq C h^{-\lambda} \left( \frac{|z-y|}{h} \right)^{1-\lambda} \|x\|_{L^\infty}$$

$$\leq C h^{-\lambda} \|x\|_{L^\infty}.$$

Now we turn over to the estimation of  $\sigma_h$ . Referring for details to Ciarlet pp.

there exists to any

$\Delta \in \mathcal{T}_h$  a set of points  $\{P_j = P_j^0 \mid j=1, \dots, J\}$

( $J = \dim \mathcal{P}_{m-1}$ , the space of polynomials of degree less than  $m$ ) with the following properties:

(i) the conditions

$$(5.27) \quad p^0(P_j) = r_j \quad \text{for } j=1, \dots, J$$

define uniquely a polynomial  $p^0$  of degree less than  $m$ .

(ii) if  $r_j = r_j^0$  coincide with the values in  $P_j^0$  of a function  $v$  continuous in  $\Omega$  then the function  $x$  defined by

$$(5.28) \quad x_{1\delta} = p^0$$

is continuous in  $\Omega$ .

Now let  $p$  be the restriction to a  $\Delta \in \mathcal{T}_h$  fixed of the interpolation of a function  $v \in C^{0,\lambda}$ . For convenience let - possibly after a translation - the origin coincide with one of the corners of  $\Delta$ , say  $P_1$ . Then  $p$  has the structure

$$(5.29) \quad p(x) = \sum_{|\beta| < m} x^\beta c_\beta(v) h^{-|\beta|}$$

with

$$(5.30) \quad x^\beta = x_1^{\beta_1} \cdots x_N^{\beta_N}$$

and

$$(5.31) \quad c_\beta(v) = \sum_{j=1}^J c_\beta^j v(P_j).$$

The  $K$ -regularity of the subdivision  $\mathcal{T}_h$  leads to the uniform boundedness of the  $c_\beta^j$  independent of  $h$ .

Since the function  $v = 1$  is reproduced by the interpolation we have

$$(5.32) \quad \sum_{j=1}^J c_\beta^j = \begin{cases} 1 & \text{for } |\beta|=0, \\ 0 & \text{for } |\beta| \geq 1. \end{cases}$$

This gives on the one hand

$$(5.33) \quad c_0(v) = v(P_1)$$

and on the other hand for  $c_\beta$  with  $|\beta| \geq 1$  a representation

$$(5.34) \quad c_\beta(v) = \sum_{j_1, j_2} \tilde{c}_\beta^{j_1, j_2} (v(P_{j_1}) - v(P_{j_2}))$$

with some  $\tilde{c}_\beta^{j_1, j_2}$  also uniformly bounded. With  $x \in \Delta$  we get

$$(5.35) \quad v(x) - p(x) = v(x) - v(P_1) - \sum_{1 \leq |\beta| < m} \frac{x^\beta}{\beta!} c_\beta$$

For  $v \in C^{0,\lambda}$  we have

$$(5.36) \quad |v(x) - v(P_1)| \leq [v]_\lambda h^\lambda.$$

Because of  $|x| \leq h$  in  $\Delta$  we get with (5.34)

$$(5.37) \quad \left| \sum_{1 \leq |\beta| < m} \{\dots\} \right| \leq C \max |v(P_{j_1}) - v(P_{j_2})| \\ \leq C [v]_\lambda h^\lambda.$$

This proves the first part of the approximation property

$$(5.9) \text{ with } \delta_h \leq C h^\lambda.$$

In order to prove the second part we consider firstly two points  $x, y$  contained in one of the simplices  $\Delta$ . Then with  $d = |x-y|$  we have  $d \leq h$  and

$$(5.38) \quad p(x) - p(y) = \sum_{1 \leq |\xi| \leq m} h^{-|\xi|} (x^\xi - y^\xi) c_\xi(v)$$

Because of

$$(5.39) \quad |x^\xi - y^\xi| \leq d h^{|\xi|-1}$$

and

$$(5.40) \quad |c_\xi| \leq c h^{\lambda} [v]_\lambda$$

we get

$$(5.41) \quad |p(x) - p(y)| \leq c d h^{\lambda-1} [v]_\lambda \leq c |x-y|^{\lambda} [v]_\lambda.$$

In case of  $d = |x-y| \leq h$  but  $x \in \Delta_1$  and  $y \in \Delta_2$  with  $\Delta_1 \neq \Delta_2$  the segment connecting  $x$  and  $y$  intersects only a finite number of  $\Delta \in \mathcal{P}_h$  because of the  $K$ -regularity.

By estimates similar to above we get for the interpolation

$$\chi = \frac{1}{h} v \text{ also then}$$

$$(5.42) \quad |\chi(x) - \chi(y)| \leq c |x-y|^{\lambda} [v]_\lambda.$$

In case of  $d = |x-y| > h$  and  $x \in \Delta_1, y \in \Delta_2$  we select two corners  $P_x, P_y$  or  $\Delta_1, \Delta_2$ . Then we have

$$(5.43) \quad \begin{aligned} X(x) - X(y) &= (X(x) - X(P_x)) + (X(P_x) - X(P_y)) \\ &\quad + (X(P_y) - X(y)). \end{aligned}$$

According to the choice of  $P_x, P_y$  we have  $|x - P_x| \leq h$  and  $|y - P_y| \leq h$  and therefore

$$(5.44) \quad \begin{aligned} |X(x) - X(P_x)| &\leq c h^\lambda [v]_\lambda, \\ |X(y) - X(P_y)| &\leq c h^\lambda [v]_\lambda. \end{aligned}$$

Since  $\chi$  is the interpolation on  $v$  we have

$$(5.45) \quad \begin{aligned} |X(P_x) - X(P_y)| &= |v(P_x) - v(P_y)| \\ &\leq |P_x - P_y|^\lambda [v]_\lambda. \end{aligned}$$

We have  $d \geq h$  and  $|P_x - P_y| \leq d + 2h \leq 3d$ . In this way also the second part of (5.9) is proven.

## 6. General Second Order Elliptic Equations

In sections 3 and 4 we presented the  $L_\infty$ -analysis of the Ritz procedure in case of the Laplacian being the prototype of an elliptic differential operator. The same results hold in the general case with  $-\Delta$  replaced by

$$(6.1) \quad Au = -a^{ik} u_{ik} + b^i u_i + du .$$

Remark 11: Throughout this section we adopt the summation convention. Lower indices indicate differentiation with respect to the corresponding variable.

The assumptions regarding the coefficients are:

(a.1) Ellipticity: There is a constant  $q > 0$  such that for all  $x \in \bar{\Omega}$  and  $\xi \in \mathbb{R}^N$

$$(6.2) \quad a^{ik} \xi_i \xi_k \geq q \sum_1^N \xi_i^2$$

holds true.

(a.2) Regularity: The coefficients  $a^{ik}$ ,  $b^i$ , and  $d$  fulfill

$$(6.3) \quad a^{ik} \in C^{2,1}, \quad b^i \in C^{1,1}, \quad d \in C^{0,1} .$$

The letter  $\bar{q}$  is used as an upper bound of all the corresponding norms.

Remark 12: Assumption (a.2) guarantees that the coefficients of the formal adjoint operator  $A^*$  defined by

$$(6.4) \quad A^* v = - (a^{ik} v)_{,k} - (\ell^i v)_i + d v$$

fulfills also (a.2).

The weak formulation of the boundary value problem

$$(6.5) \quad \begin{aligned} Au &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

is

Find  $u \in H_1^0$  such that

$$(6.6) \quad a(u, v) = (f, v)$$

holds for all  $v \in H_1^0$ .

with  $a(\cdot, \cdot)$  defined by

$$(6.7) \quad a(v, w) = \iint \{ a^{ik} v_i w_k + \ell^i v_i w + dw v \} dx.$$

Correspondingly the (generalized) Ritz approximation  $\varphi = R_h u \in S_h^0$  is characterized by the relation

$$(6.8) \quad a(\varphi, x) = (f, x) \quad \text{for } x \in S_h^0.$$

In this generality the function  $u$  defined by (6.6) resp.  $\varphi$  defined by (6.8) may not exist or may not be unique. Therefore necessarily we assume

(a.3) Existence: The problem (6.5) resp. (6.6) possesses a unique solution for  $f$  being arbitrary.

By an argument due to Schatz there is a  $h_0 > 0$  such that for  $h \leq h_0$  the Ritz approximation  $\varphi$  (6.8) is also unique.

Now we repeat the arguments of sections 3 and 4. The counterpart of (3.8) in the form

$$(6.9) \quad \| \nabla \varphi \|_{\alpha}^2 \leq C \{ a(\varphi, \tilde{\mu}^\alpha \varphi) + \| \varphi \|_{\alpha+1}^2 \}$$

is a direct consequence of Gårding's inequality

$$(6.10) \quad a(v, v) \geq \tilde{q} \| \nabla v \|_1^2 - \lambda \| v \|_1^2$$

for any  $v \in H_1^0$  with  $\tilde{q} > 0$ ,  $\lambda$  depending only on  $q, \bar{q}$ .

Remark 13: The constants  $c$  - see the beginning of section 2 - may depend in addition on (v) the bounds  $q, \bar{q}$  of the assumptions (a.1), (a.2).

Following the lines of section 3 we get from (6.9) also now the final estimate (3.22) of step 1.

The auxiliary function  $w$  - see (3.23) - is this time defined by

$$(6.11) \quad \begin{aligned} A^* w &= \mu^{-\alpha-1} \varphi \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

The estimates leading to (3.32) are derived in the same way as before.

Since the shift theorems (4.1), (4.2) are valid with  $-A$  - the Laplacian - replaced by the operator  $A$  Lemma 9 is valid with  $-A$  replaced by  $A$  on the right hand sides. As before it remains to find bounds of the terms in (4.13).

Following the lines of section 4 we consider the case of  $N = 3$  dimensions firstly. In the general case the second assertion of Lemma 10 has to be changed by the estimate

$$(6.12) \quad \|\nabla v\|_p + \|v\|_{p+1} \leq C \left\{ \|Av\|_{p-1} + \|v\|_p \right\}.$$

The last term on the right hand side may be treated as was done in the sequence (4.26) to (4.35), the details are omitted. In this way the case of  $N = 3$  dimensions is settled.

In accordance to (6.12) the a priori estimates stated in Corollary 9 have to be modified:

Corollary 9<sup>A</sup>: Let  $v \in \overset{\circ}{H}_1 \cap H_2$  resp.  $v \in \overset{\circ}{H}_1 \cap H_3$  and in addition  $Av \in \overset{\circ}{H}_1$ . Then in weighted norms for  $\beta$  arbitrary and  $N = 2$  dimensions

$$(6.13) \quad \|\nabla^2 v\|_p \leq C \{ \|Av\|_p + \|Pv\|_{p+1} + \|v\|_{p+1} \},$$

$$(6.14) \quad \|\nabla^3 v\|_p \leq C \{ \|\nabla^2 v\|_p + \|Av\|_{p+1} + \|\nabla v\|_{p+2} + \|v\|_{p+2} \}.$$

Having these shift theorems the final proof of Lemma 7 in case of a general second order elliptic differential equation follows the lines of section 4.

We will not give all the details in order to prove Lemma 9<sup>A</sup> but concentrate ourselves on the essential point. What is needed are the counterparts of (4.38) resp. (4.40) and of (4.45). By (4.38) the square sum of the second derivatives is bounded by the square of the Laplacian modulo lower order terms and a divergence term of products of first and second derivatives. In order to get the counterparts we make use of

Lemma 16: Let  $(a^{ik})$  be a positive definite and symmetric matrix according to (6.2) and let  $(b_{ik})$  be a second order tensor. Then

$$(6.15) \quad f^2 \sum_{i,k=1}^N b_{ik}^2 \leq Q^{2k} Q^{rs} b_{ir} b_{ks}.$$

Proof: Let  $\{\lambda_i^\alpha / \alpha = 1, \dots, N\}$  be an orthonormal set of eigen-vectors of the matrix  $(a^{ik})$  and  $\{\lambda^\alpha\}$  be the corresponding set of eigen-values, i. e.

$$(6.16) \quad a^{ik} \lambda_i^\alpha = \lambda^\alpha \lambda_k^\alpha \quad \text{for } \alpha = 1, \dots, N.$$

The orthogonality conditions

$$(6.17) \quad \lambda_i^\alpha \lambda_j^\beta = \delta^{\alpha\beta}$$

give rise to

$$(6.18) \quad \sum_{\alpha} \lambda_i^\alpha \lambda_k^\alpha = \delta_{ik}$$

with  $\delta^{\alpha\beta}, \delta_{ik}$  denoting the Kronecker symbol

Remark 14: In the following the summation convention is not to be applied with respect to greek letters.

The matrix  $(a^{ik})$  admits the representation

$$(6.19) \quad a^{ik} = \sum_{\alpha} \lambda^\alpha \lambda_i^\alpha \lambda_k^\alpha .$$

Then we get

$$(6.20) \quad \begin{aligned} & a^{ik} a^{rs} b_{ir} b_{ks} = \\ & = \sum_{\alpha, \beta} \lambda^\alpha \lambda^\beta \lambda_i^\alpha \lambda_k^\alpha \lambda_r^\beta \lambda_s^\beta b_{ir} b_{ks} \\ & = \sum_{\alpha, \beta} \lambda^\alpha \lambda^\beta / \delta^{\alpha\beta} /^2 \end{aligned}$$

with

$$(6.21) \quad \tilde{\ell}^{\alpha\beta} = 2_i^{\alpha} 2_k^{\beta} b_{ik} .$$

Because of  $\lambda^{\alpha} \geq q$  we get therefore

$$(6.22) \quad \begin{aligned} q^{ik} q^{rs} b_{ir} b_{ks} &\geq q^2 \sum_{\alpha, \beta} |\tilde{\ell}^{\alpha\beta}|^2 \\ &= q^2 \sum_{\alpha, \beta} 2_i^{\alpha} 2_k^{\beta} b_{ik} 2_r^{\alpha} 2_s^{\beta} b_{rs} . \end{aligned}$$

With the help of (6.18) we come from the last inequality to

$$(6.15) .$$

Now we apply (6.15) with  $b_{ik} = v_{ik}$ . Then we get

$$(6.23) \quad q^2 \|v\|_p^2 \leq \iint \mu^{\beta} (q^{ik} q^{rs} v_{ir} v_{ks}) dx.$$

Besides of lower order terms the right hand side differs

from  $\|Av\|_p^2$  by the weighted integral of the difference

$$(6.24) \quad \begin{aligned} q^{ik} q^{rs} v_{ir} v_{ks} - (q^{ik} v_{ik}) (q^{rs} v_{rs}) &= \\ &= (q^{ik} q^{rs} v_i v_{ks})_r - (q^{ik} q^{rs} v_i v_{rs})_k \\ &\quad - (q^{ik} q^{rs})_r v_i v_{ks} + \\ &\quad + (q^{ik} q^{rs})_k v_i v_{rs} . \end{aligned}$$

This leads to an inequality of the structure

$$(6.25) \quad \begin{aligned} q^2 \|D^2 V\|_p^2 &\leq \|A V\|_p + \\ &+ e \left\{ \|D^2 V\|_p^2 \|D V\|_{p+1} + \|D V\|_{p+1}^2 + \|V\|_{p+1}^2 \right\} + \\ &+ \frac{\phi \mu^{-p} \alpha^{ik} \alpha^{rs} v_i \{v_{ks} n_r - v_{rs} n_k\}}{\partial Q} ds. \end{aligned}$$

As is to be expected in view of (4.40) the boundary integral gives after evaluation of the sums

$$(6.26) \quad \phi = \frac{\phi \mu^{-p} (\alpha^{11} \alpha^{22} - (\alpha^{12})^2)}{\partial Q} (v_y dv_x - v_x dv_y).$$

By the way (4.44) was derived in the present case we come to (6.13).

The proof of (6.14) follows the same lines. Of course the formulae become somehow lengthy but there are no additional difficulties.

Bibliography

1. CIARLET, Ph.G.  
The finite element method for elliptic problems.  
North-Holland Publishing Company, Amsterdam 1977.
2. CIARLET, Ph.G. and P.A. RAVIART  
The combined effect of curved boundaries and numerical integration in isoparametric finite element methods  
Proc. of Conf. "The mathematical foundations of the finite element method with applications to partial differential equations".  
Acad. Press 409-474 (1972).
3. MORREY, C.B.  
Multiple integrals in the calculus of variations.  
Springer Verlag, New York (1966).
4. NATTERER, F.  
Über die punktweise Konvergenz Finiter Elemente.  
Num. Math. 25, 67-77 (1975).
5. NITSCHE, J.  
Zur Konvergenz von Näherungsverfahren bezüglich verschiedener Normen.  
Num. Math. 15, 224-228 (1970).
6. NITSCHE, J.  
 $L_\infty$ -Convergence of finite element approximation.  
2. Conf. on Finite Elements, Rennes (1975).
7. SCHATZ, A.H.  
An observation concerning Ritz-Galerkin methods with indefinite bilinear forms.  
Math. Comp. 28, 959-962 (1974).
8. SCHATZ, A.H. and L.B. WAHLBIN  
Maximum norm estimates in the finite element method on plane polygonal domains.  
Math. Comp. 32, 73-105 (1978).
9. SCOTT, R.  
 $L^\infty$  estimates for the finite element method on irregular meshes.  
Math. Comp. 30, 681-697 (1976).
10. ZLAMAL, M.  
Curved elements in the finite element method  
part I: SIAM J. Numer. Anal. 10, 229-240 (1973)  
part II: SIAM J. Numer. Anal. 11, 347-362 (1974).

ANALYSIS OF SOME CONTACT PROBLEMS IN NONLINEAR ELASTICITY

J. T. ODEN

University of Texas

Summary of a Lecture presented

the

Special Year in Numerical Analysis

at the

University of Maryland

College Park, Maryland

February 24, 1981

## ANALYSIS OF SOME CONTACT PROBLEMS IN NONLINEAR ELASTICITY

J.T. ODEN

Texas Institute for Computational Mechanics  
 The University of Texas

1. INTRODUCTION

In this communication, I shall outline some results recently obtained on the analysis of certain classes of contact problems in elasticity as well as some work in progress on this subject. Complete results can be found in some forthcoming papers (e.g., [1,2]).

The Signorini problem with Coulomb friction is characterized by the variational inequality

$$a(u, v-u) + J(v, u) - J(u, u) \geq f(v-u) \quad \forall v \in K \quad (1.1)$$

where

$$\begin{aligned} a(u, v) &= \int_{\Omega} E_{ijkl} u_{k,l} v_{i,j} dx \\ J(u, v) &= \int_{\Gamma_C} v_F |\sigma_n(u)| |v_T| ds \\ f(v) &= \int_{\Omega} \tilde{f} \cdot \tilde{v} dx + \int_{\Gamma_F} t \cdot v ds \end{aligned} \quad (1.2)$$

Here the usual notations of elasticity theory are employed:  $E_{ijkl}$

are the elasticities,  $u_k, v_k$  components of admissible displacements,  $v_F$  the coefficient of friction,  $u_{k,l} = \partial u_k / \partial x_l$ ,  $f$  the body forces, and  $t$  the surface tractions. The stress tensor has components  $\sigma_{ij}(v) = E_{ijkl} u_{k,l}$  and the normal stress component on the boundary is  $\sigma_n(v) = \sigma_{ij}(v) n_i n_j$ ,

$n_i$  being the components of a unit normal to  $\Gamma$ . In (1.2),  $v_T$  denotes the tangential components of  $v$  on  $\Gamma_C$ . The body  $\Omega \subset \mathbb{R}^N$  is open and bounded with smooth boundary  $\Gamma$  and  $\Gamma$  is composed of three parts:  $\Gamma_D$  on which displacements are prescribed,  $\Gamma_F$  on which forces are prescribed, and the candidate contact area  $\Gamma_C$ . Here  $K$  is a subset of a Hilbert space  $V$ ,

$$\begin{aligned} V &= \left\{ v \in (H^1(\Omega))^N \mid v = 0 \text{ a.e. on } \Gamma_D \right\} \\ K &= \left\{ v \in V \mid v \cdot n \leq s \text{ on } \Gamma_C \right\} \end{aligned} \quad (1.3)$$

In (1.3),  $v \cdot n$  denotes the normal trace of  $v_i$  on  $\Gamma$  ( $v \cdot n = \gamma(v_i) n_i$ ),  $n$  being a unit outward normal to  $\Gamma$ ,  $v \cdot n \in H^{1/2}(\Gamma_C)$  and  $s$  is the "initial gap" between the body and the foundation. The space  $V$  is equipped with the norm,

$$\| v \|_1 = \left\{ \int_{\Omega} v_{i,j} v_{i,j} dx \right\}^{1/2} \quad (1.4)$$

and the bilinear form  $a: V \times V \rightarrow \mathbb{R}$  is symmetric,  $V$ -elliptic, and continuous;

i.e. constants  $M, m > 0$  exist such that

$$a(u, v) \leq M \|u\|_1 \|v\|_1 ; a(v, v) \geq m \|v\|_1^2 \quad (1.5)$$

for all  $u, v \in V$ .

For a more elaborate description of conventions and notations for these classes of problems, see Kikuchi and Oden [3]. My aim here is to outline some results on an analysis of (1.1) and certain alternative formulations of contact problems.

## 2. SPECIAL CASES

Minimizers of the energy functional

$$F: K \rightarrow \mathbb{R} ; F(v) = \frac{1}{2}a(v, v) - f(v) - J(v, v) \quad (2.1)$$

are also solutions of (1.1). The functional  $F$  is non-convex and non-differentiable on  $K$ . In general, no existence theory is available for this class of problems and it is felt by some mechanicians that Coulomb's law is not a good model for general frictional phenomena. Because of these mathematical and physical difficulties, alternative formulations have been sought. These take the form of special cases of (1.1) which are more mathematically tractable and on completely different formulations based on alternatives to Coulomb's law.

As special cases of (1.1), we mention:

I. The Signorini Problem - in which friction is ignored. Then we have the problem,

$$a(u, v-u) \geq f(v-u) \quad \forall v \in K \quad (2.2)$$

II. The Friction Problem with Prescribed Normal Stress. Here we set  $|\sigma_n(u)| = g \geq 0$ ,  $g \in L^\infty(\Gamma_C)$  and consider the variational inequality,

$$a(u, v-u) + j(v) - j(u) \geq \bar{f}(v-u) \quad \forall v \in K \quad (2.3)$$

where

$$j(v) = \int_{\Gamma_C} g |v_T| \, ds \quad (2.4)$$

and  $\bar{f}(v) = f(v) + \int_{\Gamma_F} F_n v \cdot n \, ds$ ,  $F_n$  being a prescribed normal force.

III. Perturbed Problems. Since  $j$  of (2.4) is non-differentiable, we introduce

$$\phi_\epsilon(v) = \begin{cases} |v_T| - \epsilon/z & \text{if } |v_T| \geq \epsilon \\ \frac{1}{2\epsilon} v_T \cdot v_T & \text{if } |v_T| < \epsilon \end{cases} \quad (2.5)$$

and

$$j_\epsilon(v) = \int_{\Gamma_C} \phi_\epsilon(v) \, ds \quad (2.6)$$

as a differentiable perturbation of  $j$ ;

$$\langle D\phi_\epsilon(u), v \rangle = \int_{\Gamma_C} g \frac{\partial \phi_\epsilon(u+\theta v)}{\partial \theta} \Big|_{\theta=0} \, ds \quad (2.7)$$

We then consider the perturbed problem

$$a(u_\varepsilon, v) + \langle D j_\varepsilon(u_\varepsilon), v \rangle = \bar{f}(v) \quad \forall v \in V \quad (2.8)$$

It is easily shown that problems (2.1), (2.4), and (2.8) have unique solutions. In the case of (2.8), a solution  $u_\varepsilon$  exists for all  $\varepsilon > 0$  and, if  $u$  is the solution of (2.4), then

$$\| u - u_\varepsilon \|_1 \leq C \sqrt{\varepsilon} \quad (2.9)$$

### 3. FINITE ELEMENT APPROXIMATIONS

If  $\{v_h\}$  is a family of finite-dimensional subspaces of  $V$  constructed using standard conforming piecewise polynomial approximations of  $v$ , and endowed with the usual interpolation properties for quasi-uniform mesh refinements, and if the solution to (2.3) is in  $(H^2(\Omega))^N \cap V$ , then one can show that

$$\| \tilde{u}_h - \tilde{u}_h^\varepsilon \|_1 \leq C_1 \sqrt{\varepsilon} \quad (3.1)$$

and

$$\| u - \tilde{u}_h \|_1 \leq C_2 h \quad (3.2)$$

where  $C_1$  and  $C_2$  are independent of  $\varepsilon$  and  $h$  and  $u_h$  and  $u_h^\varepsilon$  are solutions of the discrete problems,

$$\begin{aligned} & a(u_h^i, v_h) + \left\langle \partial f_\epsilon(u_h^i), v_h \right\rangle = f(v_h) - v_h \in V_h \\ & a(u_h, v_h - u_h) + \int_{\Gamma_C} g(|v_{h_T}| - |u_{h_T}|) ds \\ & \quad \geq I(v_h - u_h) - v_h \in K_h \end{aligned} \quad (3.3)$$

We have solved (3.3)<sub>2</sub> for a number of different choices of data, polynomial approximations  $v_h$ , and computed rates of convergence are consistent with (3.1) and (3.2).

#### 4. NON-LOCAL FRICTION

As an alternative to Coulomb's law, we consider the non-local law,

$$\left. \begin{array}{l} |\sigma_T(u)|(\mathbf{x}) \leq v_F S[\sigma_n(u)](\mathbf{x}) \Rightarrow u_T = 0 \\ |\sigma_T(u)|(\mathbf{x}) = v_F S[\sigma_n(u)](\mathbf{x}) \Rightarrow \exists \lambda > 0 \text{ s.t.} \\ u_T = -\lambda \sigma_T(u), \mathbf{x} \in \Gamma_C \end{array} \right\} \quad (4.1)$$

where  $S$  is a completely continuous map from  $H^{-1/2}(\Gamma_C)$  into  $L^2(\Gamma_C)$  such that  $t \geq 0 \Rightarrow S(t) \geq 0$ . For instance,  $S$  could be taken as an extension of the map,

$$S(\sigma_n(u)) = \int_{\Gamma} \omega_p(|\mathbf{x}-\mathbf{y}|) \sigma_n(u)(\mathbf{y}) dy \quad (4.2)$$

where

$$\omega_\rho \in C_0^\infty(\Gamma), \omega_\rho \geq 0, \omega_\rho(r) = 0 \text{ for } r \geq \rho$$

Then  $\rho \in \mathbb{R}^+$  is a material property of the contact surfaces.

The variational principle for the nonlocal problem assumes the form,

$$\begin{aligned} u \in K: a(u, v-u) + \int_{\Gamma_C} v S(\sigma_n(u)) (|v_T| - |u_T|) ds \\ \geq \bar{f}(v-u) \quad \forall v \in K \end{aligned} \quad (4.2)$$

We summarize some results on this problem due to Duvant [4]; see also Demkowicz and Oden [5]:

1.  $\forall \tau \in L^2(\Omega), \tau \geq 0, \exists$  a unique  $u_\tau \in V$  such that

$$a(u_\tau, v-u_\tau) + \int_{\Gamma_C} \tau (|v_T| - |u_{T\tau}|) ds \geq f(v-u_\tau) \quad v \in K$$

2. The correspondence  $B: L^2(\Gamma_C) \rightarrow V$  given by

$$B(\tau) = u_\tau$$

is continuous, and the normal stress

$$\sigma_n(u_\tau) = \sigma_n(B(\tau))$$

is well defined in  $H^{-1/2}(\Gamma_C)$ .

3. The map  $T: L^2(\Gamma_C) \rightarrow L^2(\Gamma_C)$  defined by the composition

$$T = vS\sigma_n(-\sigma_n) \circ B$$

is weakly sequentially continuous and has at least one fixed point.

4. Let  $\psi^* \geq 0$ ,  $\psi^* \in L^2(\Gamma_C)$  be a fixed point of  $T$ . Then it is trivial to show that the function

$$u^* = B(\psi^*)$$

is, in fact a solution to (4.2).

### 5. ALGORITHM

The above steps lead to an obvious algorithm for the numerical solution of (4.2):

1. Solve (a finite element approximation) of Signorini's problem without friction for the contact pressure  $\sigma_n(u_h)$ .
2. Set  $\tau_h^{(1)} = -\sigma_n(u_h)$  and solve  $(3.3)_1$  for  $u_h^{(1)}(u_h^{\varepsilon(1)})$  for the choice  $g = \tau_h^{(1)}$ .
3. Compute  $\tau_h^{(2)}$ ,

$$\tau_h^{(2)} = T_h(\tau_h^{(1)})$$

where  $T_h$  is a finite element approximation of  $T$ .

4. Continue this process until  $\| \tau_h^{(i)} - \tau_h^{(i-1)} \|$

is less than a preassigned tolerance.

We are coding this algorithm at present and should have results soon since steps 1 and 2 can be handled using existing codes.

ACKNOWLEDGEMENT: My work on this problem was supported by the U.S. Air Force Office of Scientific Research under contract F49620-78-C-0083.

#### REFERENCES

1. Campos, L., Oden, J. T., and Kikuchi, N., Computer Methods in Applied Mechanics and Engineering, (to appear)
2. Oden, J. T., "Analysis of a Class of Contact Problems with Friction by Finite Element Methods," (MAFELAP 1981 - Brunel University) The Mathematics of Finite Elements and their Applications, Vol. IV, Academic Press, Ltd., London, (to appear).
3. Kikuchi, N. and Oden, J. T., Contact Problems in Elasticity, SIAM Publications, Philadelphia, Pa. (to appear)
4. Duvant, G., "Equilibre d'un solide elastique avec contact et frottement de Coulomb," Compte Rendus, t. 290, Feb. 4, 1980, pp.263-265.
5. Demkowicz, L. and Oden, J. T., "On Some Contact Problems with Coulomb Friction," TICOM Rept., Austin, 1981 (in press).

SINGLE STEP METHODS FOR LINEAR DIFFERENTIAL EQUATIONS  
IN BANACH SPACES

by

Vidar Thomée

Chalmers University of Technology  
Goteborg, Sweden

SINGLE STEP METHODS FOR LINEAR DIFFERENTIAL EQUATIONS  
IN BANACH SPACES

Vidar Thomée

Chalmers University of Technology  
Goteborg, Sweden

Our purpose in this paper is to present some results obtained in Brenner and Thomée [7], [8], and Brenner, Crouzeix, and Thomée [5] for time discretization of the initial value problem

$$(1) \quad \frac{du}{dt} = Au + f(t) \quad \text{for } t \geq 0, \quad u(0) = v,$$

in a Banach space  $X$  where  $A$  is a closed linear operator which generates a strongly continuous semigroup  $E(t) = e^{tA}$  on  $X$ .

In Section 1, which is a summary of [7], we are concerned with the homogeneous equation and study approximations of the semigroup at  $t = nk$  of the form  $E_k^n = r(kA)^n$  where  $r$  is an  $A$ -acceptable rational approximation of  $e^z$ . In Section 2 we examine some consequences for time discretization of equations which are already discretized with respect to a space variable; the material in this section is not contained in the above references. In Section 3, which corresponds to [8], we discuss, with applications to hyperbolic problems in mind, some modifications in the case that  $A$  generates a group on  $X$ . In Section 4, based on [5], finally, we attend to the full inhomogeneous equation in (1).

1. Time discretization of the homogeneous equation.

Let  $X$  be a Banach space with norm  $\|\cdot\|$ . Consider the initial value problem

$$(1.1) \quad \frac{du}{dt} = Au \text{ for } t \geq 0, \quad u(0) = v,$$

which we assume correctly posed in the sense that the closed linear operator  $A$  generates a strongly continuous semigroup  $E(t) = e^{tA}$  on  $X$  with, for some  $C_0$  and  $\omega \geq 0$ ,

$$(1.2) \quad \|E(t)\| \leq C_0 e^{\omega t} \text{ for } t \geq 0.$$

We shall be interested in approximating the solution  $u(t) = E(t)v$  of (1.1) by a single step discrete method so that with  $k$  the time step,  $u(t)$  is approximated at  $t = nk$  by  $u_n$ , defined recursively by

$$u_{n+1} = E_k u_n = r(kA)u_n, \quad n = 0, 1, \dots, \quad u_0 = v,$$

where  $r(z)$  is a rational function approximating the exponential  $e^z$ . We have then

$$u_n = E_k^n v = r(kA)^n v \text{ for } n \geq 0,$$

and we shall therefore be concerned with the stability and convergence properties of the operator  $E_k$ .

We shall assume below, for  $E_k$  to be well defined for any operator  $A$  of the type considered, that  $r(z)$  has no poles in  $\operatorname{Re} z \leq 0$ . More precisely, we shall assume that  $r$  is  $A$ -acceptable, or

(i)  $|r(z)| \leq 1$  for  $\operatorname{Re} z \leq 0$ ,  
(ii)  $r(z) = e^z + o(z)$  as  $z \rightarrow 0$ .

Our first aim is to present the following result which was conjectured by R. Hersh and T. Kato [13].

Theorem 1.1. Under the above assumptions there are constants  $C_0$  and  $\kappa$  depending only on the rational function  $r$  such that for any  $A$  with (1.2),

$$\|E_k^n\| \leq C_0 C_1 n^{1/2} e^{\omega x t} \text{ for } t = nk \geq 0.$$

It is possible to show that this result is best possible. For example, with  $X = L_\infty(R)$  and  $A = d/dx$  we have  $(E(t)v)(x) = v(x+t)$  and  $\|E(t)\| = 1$ . For the Crank-Nicolson scheme defined by  $r(z) = (1 + \frac{1}{2}z)/(1 - \frac{1}{2}z)$  one may then show that

$$\|E_k^n\| \geq cn^{1/2} \text{ with } c > 0.$$

Under additional hypotheses on  $r$ , however, it is possible to show improvements of the above result. We shall assume that the  $A$ -acceptable function  $r$  satisfies the following more precise condition, namely

$$(*) \quad |r(i\xi)| < 1 \text{ for } 0 \neq \xi \in R, \text{ and } |r(\infty)| < 1;$$

there exist positive integers  $p, q$  with  $p$  even,  
 $p \geq q + 1$  and a positive number  $\gamma$  such that for  $\xi \in R$ ,

$$r(i\xi) = \exp(i\xi + \psi(\xi)) \text{ with } \psi(\xi) = O(\xi^{q+1}) \text{ as } \xi \rightarrow 0$$

and

$$\operatorname{Re} \psi(\xi) \leq -\gamma \xi^p \text{ for } |\xi| \leq 1.$$

Note that if  $r$  approximates  $e^z$  to order  $q$  so that

$$r(z) = e^z + o(z^{q+1}) = \exp(z + o(z^{q+1})) \text{ as } z \rightarrow 0,$$

then the above order estimate for  $\psi(\xi)$  near zero holds. Then also, by analyticity,  $\operatorname{Re} \psi(\xi) = \gamma_0 \xi^p (1 + o(1))$  as  $\xi \rightarrow 0$  for some  $p \geq q + 1$ . Assuming that  $|r(i\xi)| < 1$  for  $0 \neq \xi \in \mathbb{R}$  we conclude that  $\gamma_0 < 0$  and that  $p$  is even so that the above estimate for  $\operatorname{Re} \psi(\xi)$  holds for a suitable  $\gamma$ . We may refer to  $p$  as the order of dissipation.

Our improvement of Theorem 1.1 is then the following.

Theorem 1.2. For each A-acceptable function  $r$  satisfying  
(\*) there are constants  $C_1$  and  $\omega$  such that for each  $A$  with (1.1),

$$\|E_k^n\| \leq C_0 C_1 n^{\frac{1}{2}(1 - \frac{q+1}{p})} e^{\omega x t} \text{ for } t = nk \geq 0.$$

Note in particular that if (\*) holds with  $q + 1 \leq p$  we have stability, independently of  $X$  and  $A$ . For instance, for the backward Euler method we have

$$r(z) = 1/(1-z) = e^z + o(z^2) \text{ as } z \rightarrow 0,$$

so that  $q = 1$  and

$$|r(i\xi)|^2 = 1/(1+\xi^2) = e^{-\xi^2 + o(\xi^2)} \text{ as } \xi \rightarrow 0,$$

so that  $\operatorname{Re} \psi(\xi) = -\frac{1}{2} \xi^2(1+o(1))$  for small  $\xi$ , and  $p = 2 = q + 1$ .

As further examples, let us consider more generally the Padé approximants  $r_{j\ell} = P_\ell/Q_j$  of  $e^z$  with degree  $P_\ell = \ell$  and degree  $Q_j = j$ , for which

$$r_{j\ell}(z) = P_\ell(z)/Q_j(z) = e^z + o(z^{j+\ell+1}) \text{ as } z \rightarrow 0,$$

so that  $q = j + \ell$ . It is known (cf. [12] and [18]) that  $r_{j\ell}$  is A-acceptable if and only if  $0 \leq j - \ell \leq 2$ . For  $k = j$  we have  $|r_{jj}(i\xi)| = 1$  and we conclude from Theorem 1.1 for  $E_k = r_{jj}(kA)$  (if  $\omega = 0$ )

$$\|E_k^n\| \leq C_0 C_1 n^{1/2} \text{ for } n \geq 0.$$

For  $\ell = j - 1$  and  $j - 2$  we may use Theorem 1.2 to show stronger results. In fact, it is shown in [1] and [12] that in these cases

$$|Q_j(i\xi)|^2 = |P_\ell(i\xi)|^2 + q_{j\ell} \xi^{2j} \text{ with } q_{j\ell} > 0,$$

which implies that (\*) holds with  $p = 2j$ . We conclude in particular stability for  $\ell = j - 1$ , and for  $\ell = j - 2$  we have (if  $\omega = 0$ )

$$\|E_k^n\| \leq C_0 C_1 n^{1/(4j)} \text{ for } n \geq 0.$$

In applications it is sometimes convenient to use approximants with denominators of the form  $(1-\gamma z)^j$ . Rational functions of this type are the so-called restricted Padé approximants (see [16]).

$$R_j(z) = (-1)^j (1-\gamma z)^{-j} \sum_{m=0}^j L_j^{(j-m)}(\gamma^{-1})(\gamma z)^m,$$

where  $L_j$  denotes the Laguerre polynomial of degree  $j$ . With suitable choice of  $\gamma$ , these approximations are of order  $j + 1$ , are A-acceptable for  $j = 1, 2, 3$ , and 5 (cf. [15], [18]), and  $|R_j(\infty)| < 1$  for  $j = 2, 3$ , and 5. For  $j = 2$ , (\*) is satisfied with  $q = 3$ ,  $p = 4$ , and for  $j = 3$  with  $q = 4$ ,  $p = 6$ . In particular,  $E_k = R_2(kA)$  (the Calahan scheme) is stable, and the norm of  $E_k^n = R_3(kA)^n$  may grow as  $n^{1/12}$ .

We shall briefly indicate the technique of proof by sketching the proof of Theorem 1.1 for  $\omega = 0$ . The main idea is to use the possibility of representing certain functions of  $A$  as integrals of the form

$$(1.3) \quad f(A) = \int_{R_+} e^{tA} d\mu(t),$$

where  $\mu$  is a bounded measure. Once this is done we may conclude from (1.2) that

$$\|f(A)\| \leq C_0 \int d|\mu|(t),$$

so that in order to estimate the norm it remains to bound the total variation of the measure  $\mu$ .

Let  $M$  be the set of bounded measures  $\mu$  on  $R$  with  $\|\mu\| = \int_R d|\mu|(t)$  and let  $\hat{M}$  denote the set of Fourier transforms  $\hat{\mu}(\xi) = \int_R e^{it\xi} d\mu(t)$  of  $\mu \in M$  with norm  $m(\hat{\mu}) = \|\mu\|$ . Further, let  $\tilde{M}$  denote the set of Laplace transforms  $\tilde{\mu}(z) = \int_{R_+} e^{zt} d\mu(t)$  of  $\mu \in M$  with  $\text{supp } \mu \subset R_+$ . From a

lemma by Paley and Wiener (cf. [17] p. 10-11) it follows that if  $f$  is bounded and analytic for  $\operatorname{Re} z \leq 0$  and if  $f(i\xi) = \hat{\mu} \in \hat{M}$  then  $f \in \widetilde{M}$  and  $f(z) = \tilde{\mu}(z)$  for  $\operatorname{Re} z \leq 0$ . For such a function  $f$  we then have the representation (1.3) as is shown in Hille-Phillips [14].

It follows in particular from the above that

$$\|E_k^n\| = \|r(kA)^n\| \leq C_0 m(r(ik\xi)^n),$$

and since it is easy to see that an affine transformation does not change  $m(f)$  we have

$$\|E_k^n\| \leq C_0 m(r(i\xi)^n).$$

In order to estimate the latter quantity we need the following inequality by Carlson [10]: If  $f, f' \in L_2(\mathbb{R})$  then  $\hat{f} \in L_1(\mathbb{R})$  and

$$\|\hat{f}\|_1 \leq 2\sqrt{2}\pi \|f\|_2^{1/2} \|f'\|_2^{1/2}.$$

Since  $f$  is the inverse Fourier transform of  $\hat{f}$  we conclude

$$m(f) \leq \frac{1}{2\pi} \|\hat{f}\|_1 \leq \sqrt{2} \|f\|_2^{1/2} \|f'\|_2^{1/2}.$$

In order to show our desired estimates we shall not apply this inequality directly to  $r(i\xi)^n$ . Instead, we first introduce a partition of unity: Let  $\phi \in C_0^\infty(\mathbb{R})$  with  $\operatorname{supp} \phi \subset \{\xi; \frac{1}{2} < |\xi| < 2\}$  and

$$\sum_{j=1}^{\infty} \phi(2^{-j}\xi) = 1 \quad \text{for } |\xi| > 2.$$

Set

$$\phi_j(\xi) = \phi(2^{-j}\xi) \quad \text{for } j > 0,$$

$$\phi_0(\xi) = 1 - \sum_{j=1}^{\infty} \phi(2^{-j}\xi).$$

Then with  $r_\infty = r(\infty)$ ,

$$m(r(i\xi)^n) \leq m(r_\infty^n) + \sum_{j=0}^{\infty} m(\phi_j(r(i\xi)^n - r_\infty^n)).$$

The first term is directly seen to be bounded by 1 and it remains to estimate the general term in the sum. Since

$$|r(i\xi) - r_\infty| \leq \frac{C}{1+|\xi|} \quad \text{and} \quad |r(i\xi)| \leq 1,$$

we have

$$|r(i\xi)^n - r_\infty^n| \leq C \min(1, \frac{n}{1+|\xi|})$$

and hence

$$\|\phi_j(r(i\xi)^n - r_\infty^n)\| \leq C \min(2^{j/2}, n 2^{-j/2}).$$

Similarly

$$|\frac{d}{d\xi} r(i\xi)| \leq \frac{C}{1+|\xi|^2}$$

implies

$$\|\frac{d}{d\xi} (\phi_j(r(i\xi)^n - r_\infty^n))\| \leq C(2^{-j/2} + n 2^{-3j/2}),$$

and hence

$$m(\phi_j(r(i\xi)^n - r_\infty^n)) \leq C n^{1/2} 2^{-j/2}.$$

It follows that

$$r(i\xi)^n \leq 1 + Cn^{1/2} \sum_{j=0}^{\infty} 2^{-j/2} \leq Cn^{1/2},$$

which completes the proof.

We shall now consider the convergence of  $u_n = r(kA)^n v = E_k^n v$  to  $u(t) = E(t)v$  as  $k = t/n$  tends to zero, in the same general circumstances as above. We have

Theorem 1.3. For each A-acceptable rational approximation  $r(z)$  of  $e^z$  of order q there are constants  $C_1$  and  $\kappa$  such that for any A satisfying (1.2),

$$\|E_k^n v - E(t)v\| \leq C_0 C_1 t k^q e^{\omega \kappa t} \|A^{q+1} v\| \quad \text{for } t = nk, v \in D(A^{q+1}).$$

Notice in particular that there is no loss of accuracy in the case of non-stability.

The proof of this result consists in noting that with

$$f_{kn}(z) = z^{-q-1} (r(kz)^n - e^{tz})$$

we have

$$\begin{aligned} \|E_k^n v - E(t)v\| &= \|r(kA)^n v - e^{tA} v\| \\ &= \|f_{kn}(A) A^{q+1} v\| \leq C_0 m(f_{kn}(i\xi)) \|A^{q+1} v\|, \end{aligned}$$

and then estimating  $m(f_{kn}(i\xi))$  by our above methods.

For less regular data we have the following

Theorem 1.4. Under the assumptions of Theorem 1.3 there are constants  $C_1$  and  $\omega$  such that for any  $A$  with (1.2), and for  $s = 0, \dots, q+1$ ,  $s \neq \frac{1}{2}(q+1)$ ,

$$\|E_k^n v - E(t)v\| \leq C_0 C_1 t^{s-\beta(s)} e^{\omega \times t} \|A^s v\| \quad \text{for } t = nk, v \in D(A^s),$$

where

$$\beta(s) = s \frac{q}{q+1} + \min(0, \frac{s}{q+1} - \frac{1}{2}).$$

If in addition  $r$  satisfies (\*) the result holds with  $\beta(s)$  replaced by

$$\beta_*(s) = s \frac{q}{q+1} + \min(0, (s - \frac{1}{2}(q+1))(\frac{1}{q+1} - \frac{1}{p})).$$

Note in particular that in the stable case of (\*) with  $p = q+1$  we have  $\beta_*(s) = s q/(q+1)$ . For  $s = 0$  we recognize the growth factor of Theorems 1.1 and 1.2.

The above technique for estimating  $m(r(i\xi))^n$  was applied in the analysis of difference schemes in [6] (cf. also [9]).

Remark. For the case that  $A$  generates a holomorphic semigroup, sharper results than the above can be obtained (cf. the discussion in [7]).

2. Totally discrete schemes.

In application to the numerical solution of initial-boundary value problems for partial differential equations one has to consider the combined effect of discretization in space and time.

Our above convergence results will therefore, in general, have to be applied to an approximating semidiscrete problem

$$(2.1) \quad \frac{du_h}{dt} = A_h u_h \quad \text{for } t \geq 0, \quad u_h(0) = v_h,$$

depending on the small positive parameter  $h$ .

We shall consider below the application of the smooth data result of Theorem 1.3 in two such situations. In the first case, which might be encountered when one is concerned with a pure initial-value problem and the differential operator  $A$  is approximated by a finite difference operator,  $A_h^{q+1}v$  will be bounded for smooth  $v$  and the analysis is straight-forward. In the second case, which is typical in the finite element situation, the boundedness of  $A_h^{q+1}v$  cannot be taken for granted, and we will have to proceed differently.

In both cases we shall assume that the general assumptions of Section 1 are satisfied. In particular, we assume that (1.2) holds, for simplicity with  $\omega = 0$ , and that  $r(z)$  is an  $A$ -acceptable rational approximation of  $e^z$  which is accurate of order  $q$ .

The finite difference type case.

We assume here that we are given an approximation  $A_h : X \rightarrow X$  of  $A$  depending on the small positive parameter  $h$ , and subspaces,  $Y, Z$  of  $X$  with  $Y \cap Z$  dense in  $X$  such that

$$(a) \quad \|A_h v - Av\| \leq \epsilon_h \|v\|_Y \quad \forall v \in Y;$$

$$(b) \quad E(t)Y \subset Y \quad \text{and}$$

$$\|E(t)v\|_Y \leq C_2 \|v\|_Y \quad \forall v \in Y;$$

$$(c) \quad \|A_h^{q+1}v\| \leq C_3 \|v\|_Z;$$

$$(d) \quad A_h \text{ generates a bounded semigroup } E_h(t) = e^{tA_h} \text{ on } X \text{ with}$$

$$\|E_h(t)\| \leq C_4 \quad \text{for } t \geq 0.$$

With  $\epsilon_h$  tending to zero with  $h$ , we may think of these as consistency and stability conditions for the semidiscrete problem (2.1). We can now infer results about the completely discrete solution defined by

$$u_{kh}(t) = E_{kh}^n v = r(kA_h)^n v \quad \text{for } t = nk.$$

Theorem 2.1. Under the present assumptions we have for  $t = nk$ ,

$$\|E_{kh}^n v - E(t)v\| \leq C_2 C_4 t \epsilon_h \|v\|_Y + C_0 C_1 C_3 t k^q \|v\|_Z.$$

Proof. We have for the error between the semidiscrete and continuous problems

$$E_h(t)v - E(t)v = \int_0^t E_h(t-s)(A_h - A)E(s)v \, ds,$$

so that for  $v \in Y$ , by (d), (a), and (b),

$$\|E_{kh}^n(t)v - E(t)v\| \leq C_2 \|v\|_Z$$

On the other hand, for the error between the completely discrete and semidiscrete solutions, we have by Theorem 1.3 and (c),

$$\|E_{kh}^n v - E_h(t)v\| \leq C_0 C_1 t k^q \|A_h^{q+1} v\| \leq C_0 C_1 C_3 t k^q \|v\|_Z.$$

Together these estimates show our result.

Note that in the present situation our error estimate is of order  $O(\epsilon_h^{-1} + k^q)$  for  $v$  sufficiently smooth, even without assuming the discrete operator  $E_{kh}$  to be stable in  $X$ .

#### The finite element type case.

Here we shall assume that we are given a family of subspaces  $X_h$  of  $X$ , depending on the small positive parameter  $h$ , and for each  $h$  a projection operator  $P_h : X \rightarrow X_h$  and a semigroup  $E_h(t)$  on  $X_h$  which is known to approximate  $E(t) = e^{tA}$  in the sense that with  $Y$  a subspace of  $X$  such that  $D(A^{q+1}) \cap Y$  is dense in  $X$ , we have

$$\|E_h(t)P_h v - E(t)v\| \leq \epsilon_h(1+\gamma t) \|v\|_Y \quad \forall v \in Y.$$

Note in particular that for  $t = 0$  this shows that  $X_h$  approximates  $X$ , or more precisely,

$$\|P_h v - v\| \leq \epsilon_h \|v\|_Y \quad \forall v \in Y.$$

With  $A_h$  the generator of  $E_h(t)$  our assumptions mean that  $u_h(t) = E_h(t)v_h$  is the solution of the "semi-discrete" problem (2.1), which is now posed in  $X_h$ , and that with  $v_h = P_h v$ ,

$$\|u_h(t) - u(t)\| \leq \epsilon_h(1+\gamma t) \|v\|_Y \quad \forall v \in Y.$$

Defining in this case the completely discrete solution at  $t = nk$  by  $u_{kh}(t) = E_{kh}^n v_h$  where  $E_{kh} = r(kA_h)$  and  $v_h = P_h v$  we have now

Theorem 2.2. Under the present assumptions we have for  $t = nk$ ,

$$\|E_{kh}^n P_h v - E(t)v\| \leq \epsilon_n(\rho_n + \rho_{n-1}\gamma t) \|v\|_Y + C_0 C_1 t k^q \|A^{q+1} v\| \quad \forall v \in D(A^{q+1}) \cap Y,$$

$$\text{where } \rho_n = m(r(i\xi))^n.$$

Proof. Consider the representation

$$r(kz)^n = \int_{R_+} e^{tz} d\mu_{kn}(t).$$

Then

$$E_{kh}^n P_h v - E_k^n v = r(kA_h)^n P_h v - r(kA)^n v = \int_{R_+} (E_h(t)P_h - E(t))v d\mu_{kn}(t),$$

and hence for  $v \in Y$ ,

$$\|E_{kh}^n P_h v - E_k^n v\| \leq \epsilon_h \|v\|_Y \left( \int_{R_+} d|\mu_{kh}| + \gamma \int_{R_+} t d|\mu_{kn}| \right).$$

Here

$$\int_{R_+} d|\mu_{kn}| = m(r(i\xi))^n = \rho_n,$$

and since

$$\int_{R_+} e^{tz} t d\mu_{kn}(t) = \frac{d}{dz} (r(kz)^n) = nkr(kz)^{n-1},$$

we have

$$\int_{R_+} t d|\mu_{kn}(t)| = nkm(r(i\xi)^{n-1}) = t\rho_{n-1},$$

so that

$$\|E_{kh}^n P_h v - E_k^n v\| \leq \epsilon_h (\rho_n + \rho_{n-1} \gamma t) \|v\|_Y.$$

Since by Theorem 1.3,

$$\|E_k^n v - E(t)v\| \leq C_0 C_1 t k^q \|A^{q+1} v\|,$$

the proof is complete.

For example, for  $r$  corresponding to the backward Euler or Calahan methods,  $\rho_n$  is bounded and the convergence rate is then  $O(\epsilon_h + k^q)$  as  $k, h \rightarrow 0$ . For the Crank-Nicolson method, the above result only shows a convergence rate of  $O(\epsilon_h^{k^{-1/2}} + k^q)$  as  $k, h \rightarrow 0$  since in this case  $\rho_n = O(n^{1/2}) = O(k^{-1/2})$  for fixed  $t$ .

We shall see now, however, that even when  $\rho_n$  is unbounded it is always possible to attain a convergence rate of  $O(\epsilon_h + k^q)$  by a suitable choice of discrete initial-values, provided the given initial data are sufficiently regular. For this purpose we first define another projection  $Q_h : X \rightarrow X_h$  by

$$Q_h = (I - A_h)^{-q-1} P_h (I - A)^{q+1}.$$

Lemma 2.1. With  $Z = \{v; v \in D(A^{q+1}), (I - A)^{q+1} v \in Y\}$  we have

$$\|(Q_h - I)v\| \leq C\epsilon_h \|(I - A)^{q+1} v\|_Y \quad \forall v \in Z.$$

Proof. From

$$(1-z)^{-q-1} = \int_0^\infty \frac{t^q}{q!} e^{(z-1)t} dt$$

we conclude

$$\begin{aligned} Q_h - I &= [(I-A_h)^{-q-1} P_h - (I-A)^{-q-1}] (I-A)^{q+1} \\ &\approx \int_0^\infty \frac{t^q}{q!} e^{-t} (E_h(t) P_h - E(t)) (I-A)^{q+1} dt, \end{aligned}$$

and hence for  $v \in Z$ ,

$$\|(Q_h - I)v\| \leq \int_0^\infty \frac{t^q}{q!} e^{-t} (1 + \gamma t) dt \cdot \epsilon_h \| (I-A)^{q+1} v \|_Y,$$

which shows the lemma.

We are now ready to prove

Theorem 2.3. Under the present assumptions, and with  $P_h$  uniformly bounded for small  $h$  and  $E_h(t)$  for small  $h$  and  $t \geq 0$ , we have for  $t = nk$  and  $v \in Z$ ,

$$\begin{aligned} &\|E_{kh}^n Q_h v - E(t)v\| \\ &\leq C\epsilon_h \{ \| (I-A)^{q+1} v \|_Y + (1 + \gamma t) \| v \|_Y \} + Ctk^q \| (I-A)^{q+1} v \|_Y. \end{aligned}$$

Proof. We have by Theorem 1.3,

$$\begin{aligned} &\|E_{kh}^n Q_h v - E_h(t) Q_h v\| \leq C_0 C_1 t k^q \| A_h^{q+1} Q_h v \| \\ &\leq C_0 C_1 t k^q \| (A_h(I-A_h)^{-1})^{q+1} P_h (I-A)^{q+1} v \| \leq C t k^q \| (I-A)^{q+1} v \|_Y. \end{aligned}$$

where we have used in the last step the boundedness of  $P_h$  and of  $A_h(I-A_h)^{-1}$  the latter of which follows from the boundedness of  $E_h(t)$  and

$$A_h(I-A_h)^{-1} = -I + (I-A_h)^{-1} = -I + \int_0^\infty E_h(t)e^{-t} dt.$$

Further, using the boundedness of  $E_h(t)$  once more, and our approximation assumptions and Lemma 2.1,

$$\begin{aligned} \|E_h(t)Q_h v - E(t)v\| &\leq \|E_h(t)(Q_h - P_h)v\| + \|E_h(t)P_h v - E(t)v\| \\ &< C\{\|(Q_h - I)v\| + \|(P_h - I)v\|\} + \varepsilon_h(1+\gamma t)\|v\|_Y \\ &\leq C\varepsilon_h\{\|(I-A)^{q+1}v\| + (1+\gamma t)\|v\|_Y\}, \end{aligned}$$

which completes the proof.

In the case that  $E_{kh}$  is stable, which can happen in specific cases even with  $\rho_n$  unbounded, we may of course also choose other discrete initial data, so that for instance

$$\|E_{kh}^n P_h v - E(t)v\| = O(\varepsilon_h^{q+1}) \quad \forall v \in Z.$$

For then

$$\begin{aligned} \|E_{kh}^n (P_h - Q_h)v\| &\leq C(\|P_h - I)v\| + \|(Q_h - I)v\| \\ &\leq C\varepsilon_h(\|v\|_Y + \|(I-A)^{q+1}v\|_Y). \end{aligned}$$

### 3. Discretization of reversible initial-value problems.

In Section 1 we studied a correctly posed initial-value problem

$$\frac{du}{dt} = Au \text{ for } t \geq 0, \quad u(0) = v,$$

which was assumed correctly posed in the sense that  $A$  generates a strongly continuous semigroup  $E(t)$ . We then considered discrete approximations of this problem at  $t = nk$  of the form

$$u_n = E_k^n v = r(kA)^n v,$$

where  $r(z)$  is a rational function with  $|r(z)| \leq 1$  for  $\operatorname{Re} z \leq 0$ .

When applied to hyperbolic problems, for instance to a first order symmetric hyperbolic system in  $\mathbb{R}^d$ ,  $d \geq 2$ , which corresponds to  $A = \sum_{j=1}^d A_j \partial/\partial x_j$  (with  $A_j$  hermitian matrices) the assumptions made in Section 1 are not entirely natural. On one hand, considered as an operator in  $L_2$ , say, this  $A$  has its spectrum on the imaginary axis and it should therefore suffice to assume  $|r(z)| \leq 1$  on  $\operatorname{Re} z = 0$ . On the other hand, although in this particular instance  $A$  generates not only a semigroup but a group of operators on  $L_2$ , our results do not permit estimates in  $L_p$  for  $p \neq 2$ , since  $\sum_j A_j \partial/\partial x_j$  does not generate a semigroup on such spaces unless the  $A_j$  commute (cf. [3]). It is known, however, (cf. [4]), that the problem is now well-posed from  $W_p^s$  into  $L_p$  if  $s > d|\frac{1}{2} - \frac{1}{p}|$ .

Our purpose in this section is to extend the results of Section 1 to a situation which takes into account the above remarks. It then appears natural to assume that  $X$ ,  $X_0$ , and  $X_1$  are Banach spaces, with  $X \cap X_1$  dense in  $X_1$ , and that  $A$  generates a strongly continuous group of operators  $E(t) = e^{tA}$  on  $X$  with  $E(t)(X \cap X_1) \subset X_0$  and such that

- (i)  $\|E(t)v\|_X \leq C e^{\omega|t|} \|v\|_X$  for  $t \in \mathbb{R}$ ,  $v \in X$ ;
- (ii)  $\|E(t)v\|_{X_0} \leq C_0 e^{\omega|t|} \|v\|_{X_1}$  for  $t \in \mathbb{R}$ ,  $v \in X \cap X_1$ .

(In the above discussion we would set  $X = L_2$ ,  $X_0 = L_p$  and  $X_1 = \mathbb{W}_p^S$ .)

We shall first consider the general case of a rational function  $r$  mapping the imaginary axis into the unit disc. Such a function, satisfying in addition  $r(z) = e^z + o(z)$  as  $z \rightarrow 0$  is referred to as I-acceptable in [16]. We shall then consider an I-acceptable function  $r$  satisfying the more precise assumption (\*) of Section 1.

We then have the following stability result.

Theorem 3.1. Let  $r$  be a rational function such that  $|r(i\xi)| \leq 1$  for  $\xi \in \mathbb{R}$ . Then there are constants  $C_1$ ,  $\kappa$ , and  $k_0$  such that for  $A$  with  $E(t) = e^{tA}$  satisfying (i) and (ii), and  $L_k = r(kA)$ ,

$$\|E_k^n v\|_{X_0} \leq C_0 C_1 e^{\omega \kappa t n^{1/2}} \|v\|_{X_1}, \quad \text{for } t = nk \geq 0, \quad k \leq k_0.$$

If in addition  $r$  satisfies (\*) we have

$$\|E_k^n v\|_{X_0} \leq C_0 C_1 e^{\omega \kappa t n^{1/2}} \frac{1}{n^{\frac{1}{2}(1-L_k+1)}} \|v\|_{X_1}, \quad \text{for } t = nk \geq 0, \quad k \leq k_0.$$

The proofs of these results use the techniques of Section 1. In the present case, if  $|r(z)| \leq 1$  for  $\operatorname{Re} z = 0$ , it is possible to factor  $r$  into  $r(z) = r_-(z)r_+(z)$  with  $|r(z)| \leq 1$  for  $\operatorname{Re} z \leq 0$  and  $|r_+(z)| \leq 1$  for  $\operatorname{Re} z \geq 0$  (or  $|r_+(-z)| \leq 1$  for  $\operatorname{Re} z \leq 0$ ) and we may write  $r(A) = r_-(A)r_+(-(-A))$ . Noting that now both  $A$  and  $-A$  generate bounded semigroups (if  $\omega = 0$ ), one may again show a representation, now with integration over all of  $\mathbb{R}$ , of the form

$$r(kA)^n = \int_{\mathbb{R}} E(t)d\mu_{nk}(t),$$

with  $\mu_{nk}$  the convolution of the measures associated with the factors  $r_-$  and  $r_+$  of  $r$  and the estimate

$$\|r(kA)^n v\|_{X_0} \leq C_0 \int_{\mathbb{R}} d|\mu_{nk}(t)| \|v\|_{X_1} = C_0 m(r(i\xi)^n) \|v\|_{X_1},$$

from which the analysis proceeds as before.

One may also prove the following convergence result.

Theorem 3.2. For each I-acceptable rational approximation  $r$  of  $e^z$  of order  $q$  there are constants  $C_1$ ,  $\kappa$ , and  $k_0$ , such that for  $A$  with  $E(t) = e^{tA}$  satisfying (i) and (ii), and for  $s = 0, 1, \dots, q+1$ ,  $s \neq \frac{1}{2}(q+1)$ , we have for  $v \in D(A^s)$  with  $A^s v \in X_1$ ,

$$\|E_k^n v - E(t)v\|_{X_0} \leq C_0 C_1 t^{s-\beta(s)} k^{\beta(s)} e^{\omega \kappa t} \|A^s v\|_{X_1} \quad \text{for } t = nk, k \leq k_0,$$

where

$$\beta(s) = s \frac{q}{q+1} + \min(0, \frac{s}{q+1} - \frac{1}{2}).$$

If in addition  $r$  satisfies (\*) the result holds with  $\beta(s)$  replaced by

$$\beta_*(s) = s \frac{q}{q+1} + \min(0, (s - \frac{1}{2}(q+1))(\frac{1}{q+1} - \frac{1}{p})).$$

Note that  $\beta(q+1) = q$  so that the convergence rate is always of optimal order  $O(k^q)$  for appropriately regular initial data.

As an example we finally consider the following special approximation of  $e^z$  which was proposed by Baker and Bramble [2] and further studied in Nørsett and Wanner [16], namely, for  $m$  a positive integer

$$r(z) = P_m(z)/(1-\gamma^2 z^2)^m, \quad \gamma > 0,$$

with

$$P_m(z) = \sum_{j=0}^{2m} z^{2m-j} \sum_{\ell=[\frac{1}{2}(j+1)]}^m {}^m_{\ell} \frac{(-\gamma^2)^{m-\ell}}{(2\ell-j)!} = \sum_{j=0}^{2m} P_{mj}(\gamma) z^{2m-j}.$$

This function is accurate of order  $2m$  and I-acceptable for suitable choice of  $\gamma$ , at least for  $\gamma \geq \gamma_m$ , the largest zero of  $P_{m0}(\gamma)$ . The proof of this latter fact (Theorem 15 of [16]) is easily modified to yield that for such  $\gamma$ ,  $r$  satisfies (\*) with  $p = 2m+2$ ,  $q = 2m$ . Thus, if  $\gamma \geq \gamma_m$ , we have for  $A$  satisfying (i) and (ii),

$$\|f_n^m v\|_{X_L} \leq C_{n,m}^{-1} e^{\omega n t} n^{\frac{1}{2}(m+1)} \|v\|_{X_1}, \quad t = nk \geq 0, \quad k \leq k_0,$$

and the estimate of Theorem 3.2 holds with

$$\beta_*(s) = s \frac{q}{q+1} + \min(0, (s - \frac{2m+1}{2})(\frac{1}{2m+1} - \frac{1}{2m+2})).$$

4. Discretization of the inhomogeneous equation.

Let again  $X$  be a Banach space and assume that  $A$  generates a bounded semigroup  $E(t) = e^{tA}$  on  $X$ . Consider now the problem

$$(4.1) \quad \frac{du}{dt} = Au + f \quad \text{for } t \geq 0, \quad u(0) = v,$$

where  $f = f(t)$ . Let  $r, \tau_1, \dots, \tau_r$  be rational functions which are bounded for  $\operatorname{Re} z \sim 0$ , and define an approximate solution of (4.1) by  $u_0 = v$  and for  $n \geq 0$ ,

$$(4.2) \quad u_{n+1} = r(kA)u_n + k \sum_{j=1}^r \tau_j(kA)(t_n + k\tau_j) = E_k u_n + k(Q_k f)(t_n),$$

where  $t_n = nk$  and  $\{\tau_j\}_{j=1}^m$  are distinct numbers which for simplicity we assume in  $[0,1]$ .

We shall say that the scheme thus introduced is of order  $p$  if for all  $f$  and  $v$  which result in sufficiently regular solutions of (4.1) and with  $A$  arbitrary,

$$\rho_n = \rho_n(k; A) = u(t_{n+1}) - E_k u(t_n) - k(Q_k f)(t_n) = O(k^{p+1}) \quad \text{as } k \rightarrow 0,$$

that is, if the solution of (4.1) satisfies (4.2) with an error  $O(k^{p+1})$ .

We observe that the global error  $e_n = u(t_n) - u_n$  satisfies

$$e_{n+1} = E_k e_n + \rho_n.$$

Assuming that  $E_k$  is stable in  $X$  we shall therefore be able to analyze the error provided we have at our disposal the appropriate representation of  $\rho_n$ .

Noting that

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)

F/G 12/1

DEC 81 I BABUSKA, T - LIU, J OSBORN

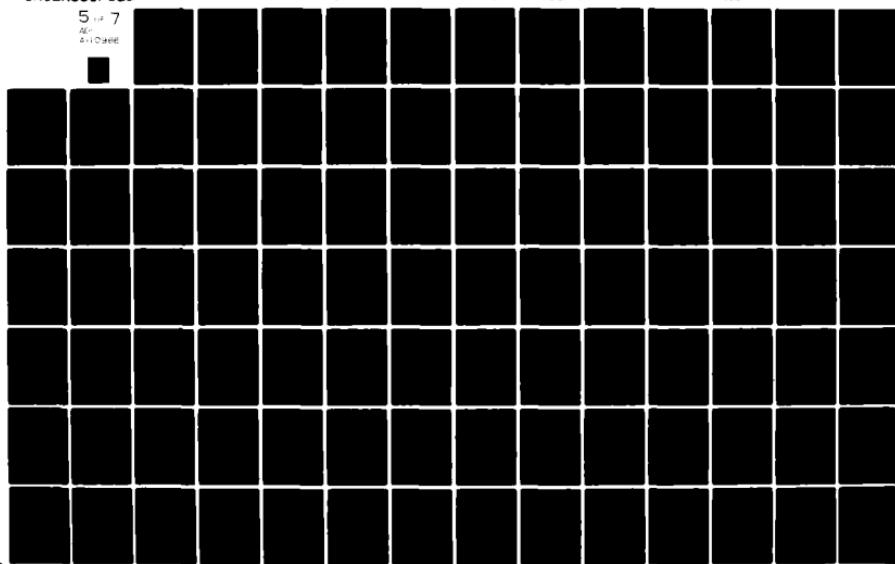
AFOSR-80-0251

UNCLASSIFIED

AFOSR-TR-82-0047

NL

5 12 7  
AFOSR  
S-10466



$$\begin{aligned} u(t_{n+1}) &= E(k)u(t_n) + k \int_0^1 E(k(1-s))f(t_n + ks)ds \\ &= E(k)u(t_n) + k(I_k f)(t_n), \end{aligned}$$

we may write

$$(4.3) \quad \rho_n(k;A) = (E(k) - E_k)u(t_n) + k((I_k - Q_k)f)(t_n).$$

In order to determine conditions on the rational functions in (4.2) for the scheme to be of order  $p$  and to find a convenient representation for  $\rho_n$  we shall develop  $\rho_n$  in Taylor series with respect to  $k$ . Since  $u$  and  $f$  are tied together by (4.1) we shall want to express  $\rho_n$  in terms of only one of them. We shall prefer here to use the data of (4.1) rather than its solution.

We begin by considering the case that (4.1) is a scalar ordinary differential equation, with  $A$  denoting multiplication by a complex number  $z$  with  $\operatorname{Re} z \leq 0$ . We have then the following result.

Lemma 4.1. With  $\rho_n$  defined by (4.3) we have

$$\rho_n(k;z) = k \sum_{\ell=0}^{p-1} k^\ell \gamma_\ell(kz) f^{(\ell)}(t_n) + k^{p+1} e^{tnz} \sigma(kz) v_{(p+1)} + \sum_{j=1}^3 R_{n,j}(k;z)$$

where

$$v_{(p+1)} = u^{(p+1)}(0) = z^{p+1} v + \sum_{\ell=0}^p z^{p-\ell} f^{(\ell)}(0),$$

$$\sigma(z) = z^{-(p+1)}(e^z - r(z)),$$

$$v(z) = z^{-(\ell+1)}(r(z) - \sum_{j=0}^{\ell} \frac{z^j}{j!}) - \frac{1}{\ell!} \sum_{j=1}^m \tau_j^\ell q_j(z),$$

and where

$$R_{n,1}(k;z) = k^{p+1} \sigma(kz) \{ -f^{(p)}(0) + \int_0^{t_n} (e^{(t_n-s)z} - 1) f^{(p+1)}(s) ds \},$$

$$R_{n,2}(k;z) = k \int_0^1 e^{(1-s)kz} \left( \int_0^{ks} \frac{(ks-\tau)^{p-1}}{(p-1)!} f^{(p)}(t_n+\tau) d\tau \right) ds,$$

$$R_{n,3}(k;z) = k \sum_{j=1}^m q_j(kz) \int_0^{k\tau_j} \frac{(k\tau_j-\tau)^{p-1}}{(p-1)!} f^{(p)}(t_n+\tau) d\tau.$$

For the proof one notes that in the present situation,

$$E(k) - E_k = e^{kz} - r(kz) = \sigma(kz) k^{p+1} z^{p+1}$$

and

$$u(t) = e^{tz} v + \int_0^t e^{(t-s)z} f(s) ds,$$

so that for the first term in  $\rho_n(k;z)$ ,

$$(E(k) - E_k) u(t_n) = k^{p+1} \sigma(kz) \{ e^{t_n z} z^{p+1} v + \int_0^{t_n} e^{(t_n-s)z} z^{p+1} f(s) ds \}.$$

Integration by parts  $p+1$  times in the last integral, Taylor expansions of  $I_k f$  and  $Q_k f$  with respect to  $k$ , and simple calculations then complete the proof.

We can now immediately show the necessity of the conditions of the next lemma.

Lemma 4.2. Necessary and sufficient for the scheme (4.2) to be of order  $p$  is that as  $z \rightarrow 0$ ,

$$(a) \quad r(z) = e^z + O(z^{p+1})$$

and

$$(b) \quad \gamma_\ell(z) = O(z^{p-\ell}) \quad \text{for } \ell = 0, \dots, p-1.$$

In fact, if (4.2) is of order  $p$  we find first by taking  $f = 0$  that  $\sigma(z)$  has to be bounded for small  $z$ , which shows (a), and then, since for a sufficiently smooth  $f$ ,  $R_{n,j}(k; z) = O(k^{p+1})$  for  $j = 1, 2, 3$  and small  $k$ , and since the  $f^{(\ell)}(t_n)$  are arbitrary, that (b) holds.

We now turn to the sufficiency of the conditions of Lemma 4.2. Recall that  $\tilde{M}$  is the set of Laplace-transforms of bounded measures on  $\mathbb{R}_+$ , and that for  $A$  the generator of a bounded semigroup  $E(t)$  on  $X$  and  $g = \tilde{\mu} \in \tilde{M}$ ,  $g(A)$  may be represented as

$$g(A) = \int_{\mathbb{R}_+} E(t)d\mu(t),$$

with

$$\|g(kA)\| \leq C_0 \int_{\mathbb{R}_+} d|\mu(t)| = C_0 m(g(i\xi)) \quad \text{if } \|E(t)\| \leq C_0.$$

Any rational function, bounded for  $\operatorname{Re} z \leq 0$ , belongs to  $\tilde{M}$ . In particular, if (b) holds we have  $\tilde{\gamma}_\ell = z^{-(p-\ell)} \gamma_\ell \in \tilde{M}$  so that we may write for  $f^{(\ell)}(t_n) \in D(A^{p-\ell})$ ,

$$(4.4) \quad \|\gamma_\ell(kA)f^{(\ell)}(t_n)\| = k^{p-\ell} \|\tilde{\gamma}_\ell(kA)A^{p-\ell}f^{(p)}(t_n)\| \leq Ck^{p-\ell} \|A^{p-\ell}f^{(p)}(t_n)\|$$

Since under our assumptions also  $\sigma \in \tilde{M}$  we have

$$\|E(t_n)\sigma(kA)v_{(p+1)}\| \leq C\|v_{(p+1)}\|,$$

and

$$\|R_{n,1}(k;A)\| \leq Ck^{p+1}\{\|f^{(p)}(0)\| + \int_0^{t_n} \|f^{(p+1)}(s)\| ds\},$$

and similarly

$$\|R_{n,2}(k;A)\| + \|R_{n,3}(k;A)\| \leq Ck^p \int_{t_n}^{t_{n+1}} \|f^{(p)}(s)\| ds.$$

We may thus write

$$p_n(k;A) = k^{p+1} \left\{ \sum_{\ell=0}^{p-1} \tilde{\gamma}_\ell(kA)A^{p-\ell}f^{(\ell)}(t_n) + E(t_n)\sigma(kA)v_{(p+1)} \right\} + \sum_{j=1}^3 R_{n,j}(k;A),$$

where under the appropriate regularity assumptions, each of the terms is  $O(k^{p+1})$  for small  $k$ . This shows the sufficiency of our conditions and thus completes the proof of Lemma 4.2.

The above estimates (4.4) for  $\gamma_\ell(kA)f^{(\ell)}(t)$  require that  $f^{(\ell)}(t) \in D(A^{p-\ell})$  for  $\ell = 0, \dots, p-1$ . In applications to partial differential equations this generally demands not only smoothness of  $f^{(\ell)}(t)$  but also that these functions satisfy certain boundary conditions which are not necessary in existence and regularity results for (4.1) and thus not natural to impose for  $t > 0$ . These requirements, however, disappear if the

coefficients  $\gamma_\ell$  vanish, which will happen for some methods.

To include this possibility in our considerations we say that the scheme (4.2) is strictly accurate of order  $p_0 \leq p$  if (a) holds together with

$$\gamma_\ell(z) = 0 \text{ for } \ell = 0, \dots, p_0 - 1.$$

The case  $p_0 = p$  is, of course, of particular interest in that the above restrictions then all disappear.

We may now state our first global error estimate which is a simple consequence of the above considerations.

Theorem 4.1. Assume that the scheme (4.2) is stable in  $X$ , accurate of order  $p$ , and strictly accurate of order  $p_0$ .

Then

$$\begin{aligned} \|u(t_n) - u_n\| &\leq C k^p \sum_{\ell=p_0}^{p-1} \sup_{s \leq t_n} \|A^{p-\ell} f^{(\ell)}(s)\| \\ &+ t_n \|v_{(p+1)}\| + t_n \|f^{(p)}(0)\| + \int_0^{t_n} (t_n - s) \|f^{(p+1)}(s)\| ds. \end{aligned}$$

The term  $\|v_{(p+1)}\|$  may be replaced by  $\|\sigma(kA)v_{(p+1)}\|$ .

Observe again that if the scheme is strictly of order  $p$  then the error is of optimal order  $p$  without any regularity conditions of the type  $f^{(\ell)}(t) \in D(A^{p-\ell})$  for  $t > 0$ . In our next result we shall see that this conclusion holds even if the scheme is strictly accurate only of order  $p - 1$ , if we make the additional assumption

$$(4.5) \quad x(z) = \frac{\gamma_{p-1}(z)}{r(z)-1} \quad \text{is bounded for } \operatorname{Re} z \leq 0.$$

Since  $r(z) = 1 + z + o(z)$  for small  $z$ , it follows easily that (4.5) holds if  $r(i\xi) \neq 1$  for  $0 \neq \xi \in \mathbb{R}$ , and if  $(r(z)-1)^{-1} = O(|z|)$  and  $q_j(z) = O(|z|^{-1})$  for large  $z$ .

Theorem 4.2. Assume that the scheme (4.2) is stable in  $X$ , accurate of order  $p$ , strictly accurate of order  $p - 1$ , and that (4.5) holds. Then

$$\begin{aligned} \|u(t_n) - u_n\| &\leq Ck^p \{t_n\|v_{(p+1)}\| + \|f^{(p-1)}(0)\| + t_n\|f^{(p)}(0)\| \\ &\quad + \int_0^{t_n} (t_n-s)\|f^{(p+1)}(s)\|ds\}. \end{aligned}$$

The term  $\|v_{(p+1)}\|$  may be replaced by  $\|\sigma(kA)v_{(p+1)}\|$ .

In the proof of this theorem one notes that the error  $e_n$  now contains the additional term

$$p_n = k^p \sum_{j=0}^{n-1} E_k^{n-1-j} \gamma_{p-1}(kA) f^{(p-1)}(t_j).$$

Setting  $s_{k,j} = \sum_{\ell=0}^j E_k^\ell$  one finds by partial summation

$$\begin{aligned} \sum_{j=0}^{n-1} E_k^{n-1-j} f^{(p-1)}(t_j) &= s_{k,n-1} f^{(p-1)}(0) \\ &\quad + \sum_{j=1}^{n-1} s_{k,n-1-j} (f^{(p-1)}(t_j) - f^{(p-1)}(t_{j-1})). \end{aligned}$$

If (4.5) holds we have  $\chi \in \tilde{M}$  and

$$P_n = k^p \chi(kA) \{ (E_k^n - I) f^{(p-1)}(0) + \sum_{j=1}^{n-1} (E_k^{n-j} - I) \int_{t_{j-1}}^{t_j} f^{(p)}(s) ds \},$$

from which the result easily follows.

We shall briefly consider application to the case when discretization also takes place in the space  $X$  as would be the case when finite element approximations are used in  $X$ . Thus let  $X_h$  be a family of subspaces of  $X$  and assume that for each  $h$  we are given a projection  $P_h : X \rightarrow X_h$  with

$$\|P_h v\| \leq C \|v\|.$$

Assume also that we are given a uniformly bounded family of semigroups  $E_h(t)$  on  $X_h$  which approximates  $E(t)$  in the sense that (cf. Section 2)

$$\|E_h(t)P_h v - E(t)v\| \leq \epsilon_h (1+\gamma t) \|v\|_Y.$$

With  $A_h$  the generator of  $E_h(t)$  we shall now study the semidiscrete problem in  $S_h$  defined by

$$\frac{du_h}{dt} = A_h u_h + P_h f \quad \text{for } t \geq 0, \quad u_h(0) = P_h v,$$

and its discretization with respect to time,

$$\begin{aligned} u_{h,n+1} &= r(kA_h)u_{h,n} + k \sum_{j=1}^m q_j(kA_h)P_h f(t_n + k\tau_j) \\ &= E_{kh} u_{h,n} + k(Q_{kh} P_h f)(t_n). \end{aligned}$$

We shall consider the error between the solution of the semidiscrete and completely discrete solutions. Combined with an error estimate for the semidiscrete problem this would show a complete error bound. We shall only present the discretized version of Theorem 4.2; obviously an analogue of Theorem 4.1 can be similarly obtained. We denote by  $Y_\theta$  the interpolation space  $Y_\theta = (X, Y)_{\theta, \infty}$  between our basic space  $X$  and its subspace  $Y$ .

We have then:

Theorem 4.3. In the present situation, assume that the scheme (4.2) is accurate of order  $p$ , strictly accurate of order  $p - 1$ , that (4.5) holds, and let  $E_{kh}$  be uniformly stable in  $X_h$ . Then

$$\|u_h(t_n) - u_{h,n}\|$$

$$\begin{aligned} &\leq C k^p \|t_n v_{(p+1)}\| + t_n \sum_{\ell=0}^p \|f^{(\ell)}(0)\|_{Y_{1-\ell/p}} \\ &+ \|f^{(p-1)}(0)\| + \int_0^{t_n} (t_n - s) \|f^{(p-1)}(s)\| ds \\ &+ C \epsilon_h \{\|v\|_Y + \|Av\|_Y + \sum_{\ell=0}^{p-1} \|f^{(\ell)}(0)\|_{Y_{1-\ell/p}}\} \end{aligned}$$

To sketch the proof, we note that application of Theorem 1.3 at once implies

$$\begin{aligned} \|u_h(t_n) - u_{h,n}\| &\leq C k^p \|t_n \sigma(kA_h) v_{h,(p+1)}\| \\ &+ \|P_h f^{(p-1)}(0)\| + t_n \|P_h f^{(p)}(0)\| + \int_0^{t_n} (t_n - s) \|P_h f^{(p+1)}(s)\| ds \end{aligned}$$

where

$$v_{h,(p+1)} = A_h^{p+1} P_h v + \sum_{\ell=0}^p A_h^{p-\ell} P_h f^{(\ell)}(0).$$

Since  $P_h$  is bounded, the terms containing  $f$  are bounded as stated. In order to estimate the first term on the right it suffices to bound

$$S = k^p \{ \sigma(kA_h) v_{h,(p+1)} - \sigma(kA) v_{(p+1)} \},$$

or, with  $\sigma_j(z) = z^j \sigma(z)$ ,

$$S = \sigma_p(kA_h) A_h v_h - \sigma_p(kA) A v + \sum_{\ell=0}^p k^\ell (\sigma_{p-\ell}(kA_h) P_h - \sigma_{p-\ell}(kA)) f^{(\ell)}(0).$$

This is done using a technique similar to the one used in Section 2.

We shall make some brief comments on how to construct schemes which satisfy our above assumptions. We first have the following alternative conditions for accuracy of order  $p$ .

Lemma 4.3. Let  $m < p$ . Then the approximation scheme (4.2) is of order  $p$  if and only if

$$(a) \quad r(z) = e^z + O(z^{p+1}) \quad \text{as } z \rightarrow 0,$$

$$(b)' \quad \gamma_\ell(z) = O(z^{p-\ell}) \quad \text{as } z \rightarrow 0, \quad \text{for } \ell = 0, \dots, m-1,$$

and there are constants  $b_1, \dots, b_m$  such that

$$(c) \quad \int_0^1 \varphi(\tau) d\tau = \sum_{j=1}^m b_j \varphi(\tau_j) \quad \forall \varphi \in \Pi_{p-1}.$$

Note in particular by applying (c) to  $\phi(t) = \prod_{j=1}^m (\tau - t_j)^2$

that the number  $m$  of quadrature points cannot be chosen smaller than  $p/2$ . On the other hand, given  $r(z)$  and  $\{\tau_j\}_1^m$  such that (a) and (c) hold, we may determine the  $q_j(z)$  such that (b)' is satisfied, for instance by solving the system

$$\gamma_\ell(z) = 0 \text{ for } \ell = 0, \dots, m-1,$$

which may be written

$$(b)'' \sum_{j=1}^m \tau_j^\ell q_j(z) = \frac{\ell!}{z^{\ell+1}} (r(z) - \sum_{j=0}^{\ell} \frac{z^j}{j!}), \quad \ell = 0, \dots, m-1.$$

Here the right hand side is a rational function which is regular for  $\operatorname{Re} z \leq 0$  by (a), and the matrix of the system is of Vandermonde's type and thus nonsingular. If  $2m \geq p$  we may always choose the  $\tau_j$  to be the Gauss points of order  $m$  to satisfy (c); if  $2m = p$  this is the only possible choice.

It is now natural to ask if the conditions (b)'' and (c) will in fact imply strict accuracy of order higher than  $m$ . In this regard we have the following:

Lemma 4.4. Let  $m < p$  and assume that (b)'' and (c) hold. Then the scheme (4.2) is accurate of order  $p$ , strictly of order  $m+1$  if and only if

$$(d) \quad r(z) = \frac{\sum_{\ell=0}^m z^{m-\ell} w(\ell)(1)}{\sum_{\ell=0}^m z^{m-\ell} w(\ell)(0)} \quad \text{where } w(t) = \prod_{j=1}^m (t - \tau_j).$$

Let now  $m = 2$  and  $p = 3$  and choose the quadrature rule

$$\int_0^1 \varphi(\tau) dt = \frac{1}{4}\varphi(0) + \frac{3}{4}\varphi\left(\frac{2}{3}\right) \quad \forall \varphi \in \Pi_2.$$

The equations for  $q_1$  and  $q_2$  are then

$$q_1 + q_2 = \frac{1}{z}(r(z)-1),$$

$$\frac{2}{3}q_2 = \frac{1}{z^2}(r(z)-1-z),$$

which gives the scheme

$$(I - (\frac{1}{2} + \frac{\sqrt{3}}{6})kA)^2 u_{n+1} = (I - \frac{\sqrt{3}}{3}kA - \frac{1}{6}(\sqrt{3}+1)k^2 A^2)u_n \\ + k(\frac{1}{4} - \frac{\sqrt{3}}{12}kA)f(t_n) + k(\frac{3}{4} - (\frac{1}{2} + \frac{\sqrt{3}}{4})kA)f(t_n + \frac{2}{3}k).$$

Again (4.5) holds and Theorems 4.2 and 4.3 apply.

References.

1. O. Axelsson, A class of A-stable methods, BIT 9 (1969), 185-199.
2. G.A. Baker and J.H. Bramble, Semi-discrete and single step fully discrete approximations for 2nd order hyperbolic equations. RAIRO Anal. Numér. 13 (1979), 75-100.
3. P. Brenner, The Cauchy problem for symmetric hyperbolic systems in  $L_p$ . Math. Scand. 20 (1967), 27-37.
4. P. Brenner, The Cauchy problem in  $L_p$  and  $L_{p,\alpha}$ . Ark. Mat. 11 (1973), 75-101.
5. P. Brenner, M. Crouzeix, and V. Thomée, Single step methods for inhomogeneous linear differential equations in Banach space. To appear.
6. P. Brenner and V. Thomée, Stability and convergence rates in  $L_p$  for certain difference schemes. Math. Scand. 26 (1970), 5-23.
7. P. Brenner and V. Thomée, On rational approximations of semigroups. SIAM J. Numer. Anal. 16 (1979), 683-694.
8. P. Brenner and V. Thomée, On rational approximations of groups of operators, SIAM J. Numer. Anal. 17 (1980), 119-125.
9. P. Brenner, V. Thomée, and L. Wahlbin, Besov Spaces and Applications to Difference Methods for Initial Value Problems, Lecture notes in Mathematics 434, Springer-Verlag, Berlin, 1975.
10. F. Carlson, Une inégalité, Ark. Mat. 25B (1935), 1-5.
11. M. Crouzeix, Sur l'approximation des équations-différentielles opérationnelles linéaires par des méthodes de Runge-Kutta, Thèse, Université de Paris VI, 1975.
12. B.L. Ehle, A-stable methods and Padé approximations to the exponential, SIAM J. Math. Anal. 4 (1973), 671-680.
13. R. Hersh and T. Kato, High-accuracy stable difference schemes for well-posed initial value problems. SIAM J. Numer. Anal. 16 (1979), 670-682.
14. E. Hille and R.S. Phillips, Functional Analysis and Semi-Groups, American Mathematical Society, Providence, RI, 1957.

15. S.P. Nørsett, One step methods of Hermite type for numerical integration of stiff systems, BIT 14 (1974), 63-77.
16. S.P. Nørsett and G. Wanner, The real pole sandwich for rational approximations and oscillation equations, BIT 19 (1979), 79-84.
17. R.E.A.C. Paley and N. Wiener, The Fourier Transform in the Complex Domain. AMS Colloquium Publications, New York, 1934.
18. G. Wanner, E. Hairer, and S.P. Nørsett, Order stars and stability theorems, BIT 18 (1978), 475-489.

Lars B. Wahlbin, Cornell University

First Lecture: Quasi-optimality of the  $\overset{\circ}{H}^1$  projection into finite element spaces: a brief survey with emphasis on the maximum norm.

Let  $R$  be a bounded domain in  $\mathbb{R}^N$ ,  $N \geq 2$ , with  $\partial R$  sufficiently smooth. The case of  $R$  polygonal or polyhedral will be commented on later. Let  $u$  denote a given function on  $R$ .

With  $0 < h < 1/2$  a parameter, let  $R_h = \bigcup_{i=1}^I \tau_i^h$  be mesh-domains partitioned into finite elements  $\tau_i^h$ , and assume for simplicity that  $R_h \subseteq R$ . Isoparametric modifications may be used at the boundary. We assume a quasi-uniform family of partitions; the case of a non-quasi-uniform family will be discussed later. Demand furthermore that  $\partial R_h$  is a uniformly Lipschitz family of curves with  $\text{dist}_{x \in \partial R} (x, \partial R_h) \leq Ch^2$ .

Let  $S_h$ ,  $0 < h < 1/2$ , be finite dimensional subspaces of  $W_\infty^1(R_h)$  consisting of functions  $x$  that vanish on  $\partial R_h$ , and are such that  $x|_\tau \in C^2(\bar{\tau})$ . Such functions can, after extension by zero, be regarded as belonging to  $W_\infty^1(R)$ . Typically,  $\{x|_\tau\}$  includes all polynomials of degree r-1,  $r \geq 2$  (or isoparametric modifications thereof).

Define  $u_h = P_h u \in S_h$  as the  $\overset{\circ}{H}^1$  projection of  $u$ , i.e.,

---

Lecture in the Department of Mathematics at University of Maryland,  
College Park, during their Special Year in Numerical Analysis,  
1980-81.

$$\begin{aligned}
 \int_{R_h} \nabla u_h \cdot \nabla \chi &= \int_{R_h} \nabla u \cdot \nabla \chi \\
 (1) \quad &= \sum_{i=1}^{I(h)} \left( - \int_{\tau_i^h} u \Delta \chi + \oint_{\partial \tau_i^h} u \frac{\partial \chi}{\partial n} \right), \text{ for all } \chi \in S_h.
 \end{aligned}$$

The question we consider is whether, giving a Banach space  $B$ , the following estimate holds,

$$(2) \quad \|P_h u\|_B \leq C(h) \|u\|_B,$$

with  $C(h)$  independent of  $u$ . If (2) holds it follows, upon writing  $u - P_h u = u - x - P_h(u - x)$  for  $x \in S_h$ , that

$$(3) \quad \|u - P_h u\|_B \leq (1 + C(h)) \min_{x \in S_h} \|u - x\|_B.$$

If  $C(h) = O(1)$  we call  $P_h u$  a "quasi-optimal" approximation, or  $P_h$  "stable". If  $C(h) = C_\varepsilon O(h^{-\varepsilon})$  for any  $\varepsilon > 0$ , we say that  $P_h u$  is "almost quasi-optimal", or that  $P_h$  is "almost stable".

1.  $B = \overset{\circ}{H}^1(R)$ .

(To be exact, consider the quasi-norm  $\|f\|_B = \|\nabla f\|_{L_2(R)}$ .) Here we have stability with  $C(h) = 1$ , and the factor  $(1 + C(h))$  in (3) can be replaced by 1. Similarly, we have stability in  $\overset{\circ}{H}^1(R_h)$ .

These results are trivial.

2.  $B = L_2(R)$ .

If the functions in  $S_h$  are merely continuous on  $R_h$  but not continuously differentiable, we cannot in general expect stability or

almost stability. A simple counterexample (in one dimension) is given in [2, p.58].

If  $S_h \subseteq C^1(R)$ , then (for  $u$  continuous) the boundary terms in the definition of (1) drop out so that

$$\int_{R_h} \nabla u_h \cdot \nabla x = - \sum_{i=1}^{I(h)} \int_{\tau_i^h} u \Delta x, \text{ for all } x \in S_h.$$

Hence (by density) the definition of the  $\overset{\circ}{H}^1$  projection makes sense for  $u \in L_2(R)$ . Correspondingly it was shown in [1, Theorem 6.3.8] that if  $S_h \subseteq H^2(R)$ , then  $P_h$  is stable in  $L_2(R)$ .

For  $C^0$  elements, two lines of investigation have been proposed in order to remedy the situation. In one dimension, it is well known that (for  $u$  continuous),

$$\|u - u_h\|_{L_2(R)} \leq C \min_{x \in S_h} \|u - x\|_{L_2(R)}.$$

$$x = u \text{ at meshpoints}$$

In [3] this result was extended to more general projections. In a related idea, [2], the  $L_2$ -norm is replaced by a certain mesh dependent norm. In one dimension,

$$\|f\|_{L_p^h} = (\int_R |f|^p + \sum_{x_j \text{ meshpoint}} h |f(x_j)|^p)^{1/p}, \quad 1 \leq p \leq \infty.$$

Then  $P_h$  in one dimension is stable in these norms. In more than one dimension, one knows stability for  $p = 2$ .

3.  $B = L_\infty(R_h)$ .

Note that the definition (1) makes sense for  $u$  continuous. Correspondingly one has, [12], that  $P_h$  is stable for  $r \geq 3$ , and almost stable (with  $C(h) = C \ln 1/h$ ) for  $r = 2$ .

A self-contained proof of this when  $N = 2$  and piecewise linear functions are employed will be given in my second lecture.

4.  $B = W_\infty^1(R)$ .

Assume here that  $R_h = R$ . Then  $P_h$  is stable for  $r \geq 2$ . For  $r \geq 3$ , cf. [4], [5], [6], [14], and for  $r = 2$ , see [7].

We now turn to listing some results in the maximum norm when  $\partial R$  is not smooth. The cases of  $H_2^1$  and  $L_2$  are fairly simple and left to the audience. Except partially for iv. below, the case of  $W_\infty^1$  is also left out. We start with two results for quasi-uniform meshes.

(i) Polygonal domain in  $R^2$ , quasi-uniform mesh,  $R_h = R$ . It is proved in [8] that  $P_h$  is stable in  $L_\infty$  for  $r \geq 3$ , and almost stable in  $L_\infty$  for  $r = 2$ .

(ii) Convex polyhedral domain in  $R^3$ , quasi-uniform mesh,  $R_h = R$ . By a recent result, [13], it is known that  $P_h$  is almost stable in  $L_\infty$  for  $r \geq 2$ .

We proceed to display what little is known in the maximum norm in the case of non-quasi-uniform meshes. For the case of  $L_2$  or  $H_2^1$ , cf. [1].

(iii) Polygonal domain in  $R^2$ , "orderly" refined mesh,  $R_h = R$ .

In [10] a guideline for constructing refined meshes is given. For the type of meshes considered there it is possible to show, [13], that  $P_h$  is almost stable.

(iv) Convex polygonal domain in  $R^2$ , "wilder" mesh,  $R_h = R$ .

The result described, which is not a true "stability" result, will appear in [13]. The kind of less orderly meshes considered may be of interest in connection with automatic, adaptive mesh-refinements. To describe typical conditions on the mesh family, assume for simplicity that the elements are triangular. Assume either

A: Piecewise linear functions, maximum angle condition, and

$$(4) \quad h_{\min} \geq h_{\max}^\gamma \quad \text{for some fixed } \gamma \geq 1.$$

Or assume

B: Piecewise polynomials of degree  $r \geq 3$ , minimum angle condition, and (4)

In (4),  $h_{\min}$  denotes the minimal side of any triangle in the mesh, and  $h_{\max}$  the maximal side of any triangle.

If either A or B holds, then, [13], for any  $\epsilon > 0$ ,

$$\|u - u_h\|_{L_\infty(R)} \leq C_\epsilon h_{\max}^{1-\epsilon} \min_{x \in S_h} \|u - x\|_{W_\infty^1(R)}.$$

We conclude this lecture by mentioning a local estimate in the maximum norm:

Let  $B(d)$  and  $B(2d)$  be concentric balls in  $R^N$  centered at a point in  $R$ . Assume that with a positive constant  $c$ ,  $d \geq ch_{\text{loc}}$

where  $h_{loc}$  denotes a local meshsize prevailing, in a quasi-uniform fashion, in  $B(2d) \cap R_h$ . If  $B(2d)$  intersects the boundary  $\partial R$ , assume that  $B(2d) \cap \partial R$  is a smooth curve. Then, [9],

$$\begin{aligned} & \|u - u_h\|_{L_\infty(B(d) \cap R_h)} \\ & \leq C(\ln(1/h_{loc}))^{\bar{r}} \min_{x \in S_h} \|u - x\|_{L_\infty(B(2d) \cap R_h)} \\ & + Cd^{-N/2-s} \|u - u_h\|_{H^{-s}(B(2d) \cap R_h)}. \end{aligned}$$

Here  $s > -1$ , and  $\bar{r} = 0$  for  $r \geq 3$ ,  $\bar{r} = 1$  for  $r = 2$ .

Thus, the local error in the finite element solution is estimated by two terms: The first term involves the local approximability of  $u$ . The second term takes global effects into account; however, it measures these effects in an arbitrarily "weak" way.

For an application of this result to calculation of stress intensity factors, see [11].

\*

#### References.

1. I. Babuška and A. K. Aziz, Survey lectures on the mathematical foundations of the finite element method, *The Mathematical Foundations of the Finite Element Method*, A. K. Aziz, Editor, Academic Press, New York, 1972.
2. I. Babuška and J. Osborn, Analysis of finite element methods for second order boundary value problems using mesh dependent norms, *Numer. Math.* 34, 1980, 41-62.

3. S. C. Eisenstat, R. Schreiber and M. H. Schultz, On the optimality of the Rayleigh-Ritz approximation, Research Report 83, Department of Computer Science, Yale University, 1976.
4. F. Natterer, Über die punktweise Konvergenz Finiter Elemente, Numer. Math. 25, 1975, 67-77.
5. J. A. Nitsche,  $L_\infty$ -convergence of finite element approximations, Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics 606, Springer, New York, 1977, 261-274.
6. R. Rannacher, Zur  $L^\infty$ -Konvergenz linearer Finiter Elemente, Math. Z. 149, 1976, 69-77.
7. R. Rannacher and R. Scott, A note on linear finite elements, to appear.
8. A. H. Schatz, A weak discrete maximum principle and stability of the finite element method in  $L_\infty$  on plane polygonal domains, Math. Comp. 34, 1980, 77-91.
9. A. H. Schatz and L. B. Wahlbin, Interior maximum norm estimates for finite element methods, Math. Comp. 31, 1977, 414-442.
10. A. H. Schatz and L. B. Wahlbin, Maximum norm estimates in the finite element method on plane polygonal domains, Part 2, Refinements, Math. Comp. 33, 1979, 465-492.
11. A. H. Schatz and L. B. Wahlbin, On a local asymptotic error estimate in finite elements and its use: numerical examples, Fourth International Symposium on Computer Methods for Partial Differential Equations, to appear.
12. A. H. Schatz and L. B. Wahlbin, On the quasi-optimality in  $L_\infty$  of the  $\overset{\circ}{H}^1$  projection into finite element spaces, to appear.

13. A. H. Schatz and L. B. Wahlbin, to appear
14. R. Scott, Optimal  $L^\infty$  estimates for the finite element method on irregular meshes, Math. Comp. 30, 1976, 681-697.

Lars B. Wahlbin, Cornell University

Second Lecture: The quasi-optimality in the maximum norm of the  $H^1$  projection into piecewise linear functions in the plane: a complete proof.

Orientation:

The  $\overset{\circ}{H}^1$  projection  $P_h u$  of a function  $u$  into piecewise linear functions  $S_h$  on a quasi-uniform family of triangulations of sizes  $h$  of a basic convex bounded domain  $R$  in  $R^2$  with smooth boundary satisfies

$$\|u - P_h u\|_{L_\infty(R_h)} \leq C \ln(1/h) \min_{x \in S_h} \|u - x\|_{L_\infty(R_h)}.$$

Here  $R_h \subseteq R$  is the meshdomain, and the constant  $C$  is independent of  $h$  and  $u$ .

The purpose of the present lecture is to give a self-contained proof of the result above. Note in particular that  $R_h \neq R$  and that this fact is not assumed away; a certain amount of technical detail ensues.

The Set-up:

Let  $R$  be a bounded convex domain in  $R^2$ , with smooth boundary  $\partial R$ . Let  $0 < h < 1/2$  denote a parameter, and

$$R_h = \bigcup_{i=1}^{I(h)} \tau_i^h$$

---

Lecture in the Department of Mathematics at University of Maryland, College Park, during their Special Year in Numerical Analysis, 1980-81.

a family of edge-to-edge triangulations "of  $R$ ", with  $R_h \subseteq R$ .

Assume that the family of triangulations is quasi-uniform, i.e., that there exist positive constants  $c$  and  $C$  so that for any angle in  $\tau_i^h$ ,

$$0 < c \leq \theta_i^h, \quad ch \leq \text{diameter } (\tau_i^h) \leq Ch.$$

Assume also that the boundary nodes of  $\partial R_h$  are placed on  $\partial R$ .

Then we have

$$(1) \quad \text{the number of boundary nodes is } \leq Ch^{-1},$$

$$(2) \quad \text{dist } (x, \partial R_h) \leq Ch^2, \text{ for some constant } C, \\ x \in \partial R$$

Let  $S_h$  be the space of continuous piecewise linear functions on  $R_h$ , which furthermore vanish on  $\partial R_h$ . After extension by zero, such functions can be considered as being in  $W_\infty^1(R)$ .

Let  $u$  be a given function on  $R$ , and define its  $H^1$  projection  $u_h = P_h u \in S_h$  via

$$(3) \quad \int_{R_h} \nabla u_h \cdot \nabla x = \int_{R_h} \nabla u \cdot \nabla x = \sum_{i=1}^{I(h)} \oint_{\partial \tau_i^h} u \frac{\partial x}{\partial n}, \text{ for all } x \in S_h.$$

Here Green's formula and the fact that  $\Delta x \equiv 0$  on each triangle were used.

Note that  $u_h$  is well defined for any function  $u$  which is continuous on  $\overline{R}_h$ .

#### The Main Result:

Under the hypotheses above, there exists a constant  $C$  independent of  $u$  and  $h$  such that

$$\|u - u_h\|_{L_\infty(R_h)} \leq C \ln(1/h) \min_{x \in S_h} \|u - x\|_{L_\infty(R_h)}.$$

Remark:

The main result is a special case of a more general result by A.H. Schatz and myself, see Lecture 1, point 3., in particular. In the case under consideration the proof is much simpler than in the general case, and, it is hoped, the main ideas can be discerned uncluttered by much technical detail. However, some trouble is unavoidable since  $R_h \neq R$ .

The Proof:

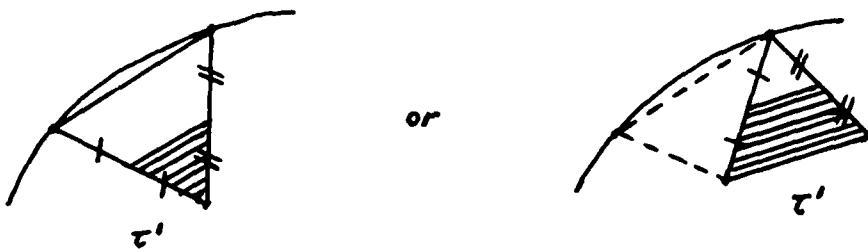
Writing  $u - u_h = (u - x) - P_h(u - x)$  for  $x \in S_h$  we see that it suffices to show

$$(4) \quad \|u_h\|_{L_\infty(R_h)} \leq C \ln(1/h) \|u\|_{L_\infty(R_h)}.$$

Let  $x_0$  be such that

$$|u_h(x_0)| = \|u_h\|_{L_\infty(R_h)}.$$

Denote by  $\tau$  a triangle such that  $x_0 \in \bar{\tau}$ . If  $\tau$  is a boundary triangle we define  $\tau'$  by "quartering" the triangle; this is described most conveniently by a figure of the two cases that can occur. ( $\tau'$  is shaded.)



If  $\tau$  is not at the boundary, let  $\tau' = \tau$ . We collect here two facts that will be used later.

(i) There is a positive constant  $c'$  such that

$$(5) \quad \text{dist}(\tau', \partial R_h) \geq c'h.$$

This is easily seen from the geometry of the situation.

(ii) There exists a constant  $C$  such that for any  $x \in S_h$ ,

$$(6) \quad \|x\|_{L_\infty(\tau)} \leq Ch^{-1} \|x\|_{L_2(\tau')}.$$

The proof of this is the same as the usual proof of an inverse property. (Equivalence of any norms on a space of polynomials of fixed degree, and scaling)

Comment.

The introduction of  $\tau'$  is made for our proof of Lemma 1 below.

We now have, by (6),

$$|u_h(x_0)| \leq Ch^{-1} \|u_h\|_{L_2(\tau')}$$

and it thus remains to show

$$(7) \quad \|u_h\|_{L_2(\tau')} \leq Ch \ln(1/h) \|u\|_{L_\infty(R_h)}.$$

A duality argument is next in store. We have

$$(8) \quad \|u_h\|_{L_2(\tau')} = \sup_{\substack{\phi \in C_0^\infty(\tau') \\ \|\phi\|_{L_2} = 1}} \int_{\tau'} u_h \phi.$$

Regard  $\phi$  as zero outside  $\tau'$ . For each such fixed  $\phi$ , let  $v$  be the solution of

$$(9) \quad -\Delta v = \phi \quad \text{in } R, \quad v = 0 \quad \text{on } \partial R.$$

We note that then by elliptic regularity,

$$(10) \quad \|v\|_{H^2(R)} \leq C.$$

Further,

$$(11) \quad v(x) = \int_{\tau'} G^x(y) \phi(y) dy$$

where

$$(12) \quad |D_x^\alpha G^x(y)| \leq \begin{cases} C(1 + |\ln|x-y||), & |\alpha| = 0, \\ C|x-y|^{-|\alpha|}, & |\alpha| = 1, 2. \end{cases}$$

Now with  $v_h = P_h v$  the  $H^1$  projection of  $v$ ,

$$(13) \quad \begin{aligned} \int_{\tau'} u_h \phi &= - \int_{R_h} u_h \Delta v = \int_{R_h} \nabla u_h \cdot \nabla v = \int_{R_h} \nabla u_h \cdot \nabla v_h \\ &= \int_{R_h} \nabla u \cdot \nabla v_h = \sum_i \oint_{\tau_i} u \frac{\partial v_h}{\partial n} = \sum_i \oint_{\tau_i} u \frac{\partial v}{\partial n} + \sum_i \oint_{\tau_i} u \frac{\partial (v_h - v)}{\partial n}. \end{aligned}$$

Here the definition of the  $H^1$  projection and Green's formula were used, and also the fact that  $\Delta v_h \equiv 0$  over each element. Note that we may assume that  $u \in C^1(\bar{R})$ , by a density argument. The first sum on the right of (13) simplifies to

$$\oint_{\partial R_h} u \frac{\partial v}{\partial n}$$

and so

$$(14) \quad \int_{\Gamma} u_h \phi = \int_{\partial R_h} \phi u - \frac{\partial v}{\partial n} + \sum_i \int_{\partial \tau_h^i} \phi u \frac{\partial (v_h - v)}{\partial n} = I_1 + I_2$$

Convention.

For the remainder of the proof we use the convention that

$$\| \cdot \|_{L_p} = \| \cdot \|_{L_p(R_h)}.$$

We next estimate the two terms in (14).

For  $I_1$ : We have

$$|I_1| \leq \|u\|_{L_\infty} \left\| \frac{\phi}{\partial R_h} \right\| \left\| \frac{\partial v}{\partial n} \right\|.$$

We use the following lemma, postponing its proof for the moment.

LEMMA 1.

$$\left\| \frac{\phi}{\partial R_h} \right\| \leq C h.$$

We thus obtain

$$(15) \quad |I_1| \leq C h \|u\|_{L_\infty}.$$

For  $I_2$ : We have

$$|I_2| \leq \|u\|_{L_\infty} \sum_i \left\| \frac{\phi}{\partial \tau_h^i} \right\| \left\| \frac{\partial (v_h - v)}{\partial n} \right\|.$$

On the unit triangle  $T$  with vertices  $(0,0)$ ,  $(1,0)$ ,  $(0,1)$  it is easily found, by use of cutoff functions and integration of derivatives in directions locally non-tangential to the boundary, that

$$\frac{1}{T} \int_T |f| \leq C (\|f\|_{L_1(T)} + \|\nabla f\|_{L_1(T)}).$$

Therefore, by scaling and by the quasi-uniformity of the mesh family, since second derivatives of  $v_h$  vanish,

$$\oint_{\Gamma_i} h \frac{\partial}{\partial n} (v_h - v) \leq C(h^{-1}) \| \nabla(v - v_h) \|_{L_1(\Gamma_i)} + \max_{|\alpha|=2} \| D^\alpha v \|_{L_1(\Gamma_i)}.$$

Hence,

$$|I_2| \leq C \| u \|_{L_\infty} \left( \max_{|\alpha|=2} \| D^\alpha v \|_{L_1} + h^{-1} \| \nabla(v - v_h) \|_{L_1} \right).$$

We now use another lemma, the proof of which will also be postponed.

LEMMA 2.

$$\max_{|\alpha|=2} \| D^\alpha v \|_{L_1(R)} \leq C h \ln(1/h).$$

Using this,

$$|I_2| \leq C \| u \|_{L_\infty} (h \ln(1/h) + h^{-1} \| \nabla(v - v_h) \|_{L_1})$$

Upon combining this estimate and (15) into (14), and (8), we find that in order to prove (7), it remains to show that

$$(16) \quad \| \nabla(v - v_h) \|_{L_1} \leq C h^2 \ln(1/h),$$

where  $-\nabla v = \phi$  in  $R$ ,  $v = 0$  on  $\partial R$ ,  $\phi \in C_0^\infty(\Gamma')$ ,  $\| \phi \|_{L_2} = 1$ .

(Of course, it also remains to verify Lemmas 1 and 2.)

Comment.

The inequality (16) can be viewed as an  $W_1^1$  estimate for a smoothed out and scaled Green's function. For higher order subspaces, one also need to estimate  $v - v_h$  in the piecewise  $W_1^2$ -norm; in our particular case this reduced to Lemma 2.

We next need to introduce some notation. Set

$$A_j = \{x : 2^{-j-1} \leq |x-x_0| \leq 2^{-j}\},$$

$$U_j = A_j \cap R_h,$$

$$d_j = 2^{-j}.$$

We have  $\overline{R_h} \subseteq \bigcup_{j=0}^{\infty} U_j$ ; the lower index is assumed to be zero for convenience. Set

$$\epsilon_* = h_h \sqrt{\bigcup_{j=0}^{\infty} U_j}$$

where with  $C_*$  a positive constant (more about it in a moment)  $\epsilon$  is defined by the requirement

$$2^{-\epsilon-1} < C_* h < 2^{-\epsilon}.$$

Note that  $\epsilon \approx \ln(1/h)$ . Let also

$$A'_j = A_{j-1} \cup A_j \cup A_{j+1}, \quad A''_j = A'_{j-1} \cup A'_j \cup A'_{j+1}, \quad \Omega'_j = A'_j \cap R_h, \quad \Omega''_j = A''_j \cap R_h.$$

Assume that  $C_*$  is so large that with a positive constant  $c$ ,

$$\text{dist}(A''_{j+1}, \cdot) \geq ch.$$

Then with a positive constant  $c$ ,

$$(17) \quad \text{dist}(\cdot_j, \cdot) \geq d_j, \quad \text{for } j=0, 1, \dots, \epsilon.$$

Comment.

We shall on occasion, Lemma 3 et seq., need to enlarge  $C_*$ .

The role of having  $C_*$  large is dual: one reason is to satisfy (17). The other is to have a certain separation between the circular boundary pieces of  $\cdot_j$ ,  $\cdot_j'$  and  $\cdot_j''$ . Basically we need that for  $j=0, 1, \dots, \epsilon$ ,

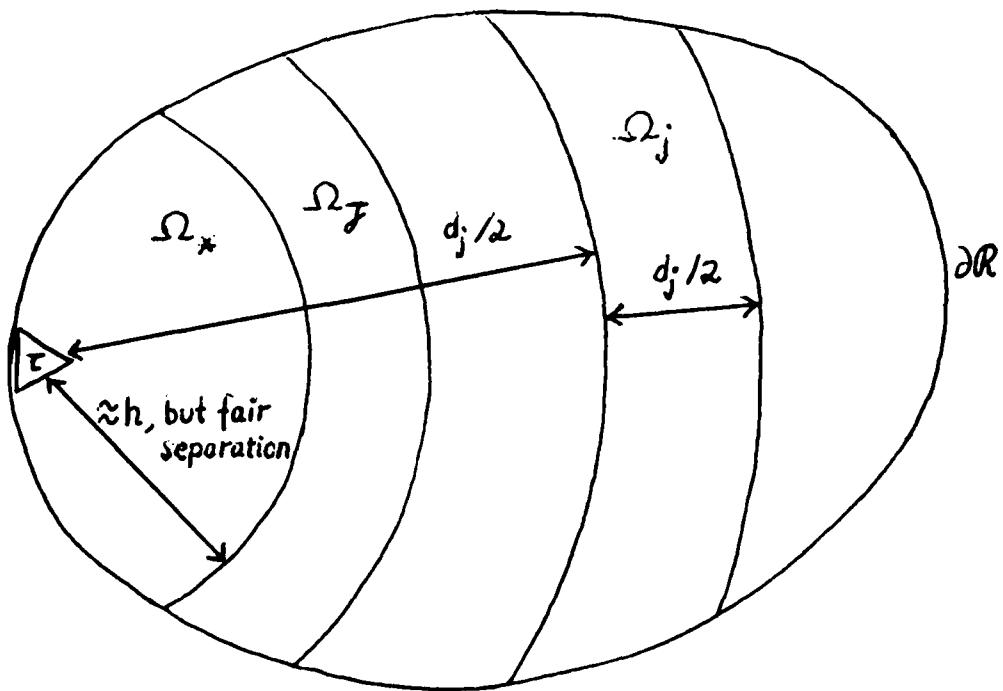
$$\Omega_j \cup \{\tau_i^h \text{ which intersect } \Omega_j\} \subseteq \Omega_j^1,$$

$$\Omega_j^1 \cup \{\tau_i^h \text{ which intersect } \Omega_j^1\} \subseteq \Omega_j^2.$$

It is realized that this will be fulfilled if  $C_*$  is large enough.

A similar state of affairs occurs in the proof of Lemma 3.

The following figure, which is not to scale, depicts the situation.



Set

$$e = v - v_h$$

and recall that we are hunting for (16). We have

$$\|v e\|_{L_1} = \|\nabla e\|_{L_1(\Omega_\star)} + \sum_{j=0}^J \|\nabla e\|_{L_1(\Omega_j)}.$$

Here by Cauchy's inequality, by a well-known result for the  $H^1$  projection, and by (10),

$$\|\nabla e\|_{L_1(\Omega_\star)} \leq C h \|e\|_{H^1(R)} \leq C h^2 \|v\|_{H^2(R)} \leq C h^2$$

so that

$$(18) \quad \|v e\|_{L_1} \leq C h^2 + \sum_{j=0}^J \|\nabla e\|_{L_1(\Omega_j)}.$$

Next, again by Cauchy's inequality,

$$(19) \quad \|\nabla e\|_{L_1(\Omega_j)} \leq C d_j \|\nabla e\|_{L_2(\Omega_j)}.$$

We shall use the following local  $H^1$ -estimate; as usual, we wait a while before giving the proof.

LEMMA 3.

Assume that  $\epsilon_\star$  is large enough. There exists a constant  $C$  such that the following holds: For  $j=0, 1, \dots, J$ ,

$$\begin{aligned} \|\nabla e\|_{L_2(\Omega_j)} &\leq C \min_{x \in S_h} (\|\nabla(v-x)\|_{L_2(\Omega'_j)} + d_j^{-1} \|v-x\|_{L_2(\Omega'_j)}) \\ &\quad + C d_j^{-1} \|e\|_{L_2(\Omega'_j)}. \end{aligned}$$

Taking  $x$  in Lemma 3 to be the interpolant of  $v$ , and remembering that  $d_j \geq ch$ , we obtain from (19) and elementary approximation theory, and using also the well known result that

$$\|e\|_{L_2(R)} \leq C h^2 \|v\|_{H^2(R)} \leq C h^2 \quad (\text{by (10)}),$$

$$\begin{aligned} \|\nabla e\|_{L_1(\Omega_j)} &\leq C d_j \|\nabla(v-x)\|_{L_2(\Omega_j')} + C \|v-x\|_{L_2(\Omega_j')} \\ &\quad + C \|e\|_{L_2(R)} \\ &\leq C d_j^2 \|\nabla(v-x)\|_{L_1(\Omega_j')} + C d_j \|\nabla(v-x)\|_{L_\infty(\Omega_j')} + C \|e\|_{L_2(R)} \\ &\leq C d_j^2 h \|v\|_{W_\infty^2(\Omega_j')} + C h^2. \end{aligned}$$

(For the last step, cf. the comment after (17).)

By the Green's function representation (11), and using (12) and (17),

$$\begin{aligned} \|v\|_{W_\infty^2(\Omega_j'')} &\leq \max_{|\tau| \leq 2} \sup_{x \in \Omega_j''} \int_{\Gamma} |D_x^\tau G^X(y)| |\tau(y)| dy \\ &\leq C d_j^{-2} \int_{\Gamma} |\tau| \leq C h d_j^{-2}. \end{aligned}$$

Therefore,

$$\|e\|_{L_1(\Omega_j)} \leq C h^2.$$

Inserting this into (18), and recalling that  $\delta \sim \ln(1/h)$ ,

$$\|ve\|_{L_1} \leq C h^2 \ln(1/h).$$

This is the desired inequality (16).

To complete the proof of our main result it remains now to verify Lemmas 1,2 and 3.

Proof of Lemma 1.

Recall that  $-\Delta v = \phi$  in  $R$ ,  $v=0$  on  $\partial R$ , where

$$\phi \in C_0^\infty(\tau'), \quad \|\phi\|_{L_2} = 1.$$

We first consider

$$\oint_{\partial R} |\nabla v| = \oint_{\partial R} \left| \frac{\partial v}{\partial n} \right| = \sup_{\substack{|\eta|_{L_\infty(\partial R)} = 1 \\ \eta \in C^\infty(\partial R)}} \oint_{\partial R} \left| \frac{\partial v}{\partial n} \right| \eta .$$

For each fixed  $\eta$ , let  $w$  denote the harmonic extension of  $\eta$  into  $R$ . Then Green's second formula and Cauchy's inequality give

$$\begin{aligned} \oint_{\partial R} \left| \frac{\partial v}{\partial n} \right| \eta &= - \int_R (\Delta v) w = \int_{\tau'} \phi w \leq C h \quad \|\phi\|_{L_2} \|w\|_{L_\infty(R)} \\ &\leq C h, \end{aligned}$$

where the maximum principle was used in the last step.

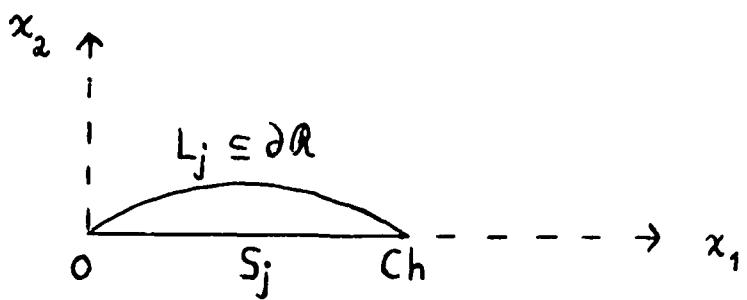
Hence,

$$(20) \quad \oint_R |\nabla v| \leq Ch.$$

We need to prove the corresponding estimate with  $\partial R$  replaced by  $\partial R_h$ . Let  $\partial R_h = \cup_j S_j$ , describing the linear segments  $S_j$  making up  $\partial R_h$ . Consider  $\sum_j \oint_{S_j} |Dv|$  where  $D$  is a generic first derivative.

Introduce (by rotation and translation) a new coordinate system locally so that

$$S_j = \{ (x_1, x_2) : 0 \leq x_1 \leq Ch, x_2 = 0 \}$$



Then  $L_j$ , the corresponding piece of  $\partial\Omega$ , is given by  
 $x_2 = b(x_1), 0 \leq x_1 \leq Ch.$

We have

$$(21) \quad (Dv)(x_1, 0) = (Dv)(x_1, b(x_1)) - \int_0^{b(x_1)} \left( \frac{\partial}{\partial x_2} Dv \right)(x_1, x_2) dx_2$$

Here, by (11) and (12), and Cauchy's inequality,

$$\begin{aligned} \left| \frac{\partial}{\partial x_2} (Dv)(x_1, x_2) \right| &= \left| \int_{\tau'}^1 \frac{\partial}{\partial x_2} D_x G^X(y) \psi(y) dy \right| \\ &\leq \frac{C}{\text{dist}(\tau', S_j)^2} \int_{\tau'}^1 \|\psi\|_{L_2} dy \leq \frac{C h \|\psi\|_{L_2}}{\text{dist}(\tau', S_j)^2} \leq \frac{C}{h}; \end{aligned}$$

in the last step we used (5). Hence from (21), since by (2) we have  
 $|b(x_1)| \leq C h^2$ ,

$$|(Dv)(x_1, 0)| \leq |(Dv)(x_1, b(x_1))| + Ch.$$

Integrating over  $S_j$  we obtain (since  $b(x_1)$  is smooth),

$$\int_{S_j} \phi |Dv| \leq C \int_{L_j} \phi |Dv| + Ch^2.$$

Summing over  $j$  gives, by (1) and (20),

$$\int_{\partial R_h} \phi |\nabla v| \leq C \int_{\partial R_h} \phi |\nabla v| + \frac{Ch^2}{h} \leq Ch.$$

This proves the Lemma. (Actually, the difference in the two integrals over  $\partial R$  and  $\partial R_h$  is  $O(h^2)$ .)

#### Proof of Lemma 2.

With  $D^2$  a generic second derivative, we estimate  $\|D^2v\|_{L_1(R)}$ . Let  $B$  denote a disc around  $x_0$  of size  $Ch$  and such that

$$\text{dist}(R \setminus B, \tau') \geq ch$$

for some positive  $c$ . We have

$$\|D^2v\|_{L_1(R)} = \int_B |D^2v| + \int_{R \setminus B} |D^2v|$$

Here, by Cauchy's inequality and (10),

$$\int_B |D^2v| \leq Ch \|v\|_{H^2(R)} \leq Ch.$$

For  $x \in R \setminus B$ , the Green's function representation (11) and (12) give, together with Cauchy's inequality,

$$|D^2v(x)| \leq \frac{C}{\text{dist}(x, \tau')^2} \quad \int_{\tau'} |\phi| \leq \frac{Ch}{\text{dist}(x, \tau')^2}.$$

By the manner in which the radius of  $B$  was chosen,

$$\int_{R \setminus B} |D^2 v| \leq Ch \int_{R \setminus B} \text{dist}(x, \tau')^{-2} dx \leq Ch \ln(1/h).$$

Hence,

$$\|D^2 v\|_{L_1(R)} \leq Ch \ln(1/h).$$

This completes the proof of the lemma.

#### Proof of Lemma 3.

We start by noting a preliminary result, often referred to as "superapproximation". Let  $\omega$  be a smooth function with  $|\omega| \leq 1$ , let  $x \in S_h$ , and let  $I \in S_h$  be the interpolant of  $\omega^2 x$ . Then for any triangle  $\tau$ , it is well known that

$$\|\omega^2 x - I\|_{H^1(\tau)} \leq Ch \max_{|\alpha|=2} \|D^\alpha (\omega^2 x)\|_{L_2(\tau)}.$$

If  $\alpha = \alpha_1 + \alpha_2$ , then since  $x$  is linear,

$$D^\alpha (\omega^2 x) = (D^{\alpha_1} \omega^2)_x + 2((D^{\alpha_1} \omega) D^{\alpha_2} x + (D^{\alpha_2} \omega) D^{\alpha_1} x).$$

Consequently,

$$(22) \quad \begin{aligned} \|\omega^2 x - I\|_{H^1(\tau)} &\leq Ch (\|\omega\|_{W_\infty^2(\tau)} \|x\|_{L_2(\tau)} \\ &+ \|\omega\|_{W_\infty^1(\tau)} \|\omega \nabla x\|_{L_2(\tau)}). \end{aligned}$$

Let now  $j$  be fixed for the rest of our argument, and set  $A = A_j$ ,  $A' = A'_j$ ;  $s = s_j$ ,  $\Omega' = \Omega'_j$ ,  $d = d_j$ . Introduce the auxiliary domains

$$A^k = \{x: \frac{d}{2} (1 - \frac{1}{2k}) \leq |x-x_0| \leq d(1 + \frac{1}{k})\}, \quad k = 2, 3, 4$$

and

$$\Omega^k = A^k \cap R_h, \text{ so that } \Omega \subseteq \Omega^4 \subseteq \Omega^3 \subseteq \Omega^2 \subseteq \Omega'$$

Consider first functions  $w_h \in S_h$  which are "discrete harmonic" in  $\mathbb{R}^2$ , i.e., such that

$$(23) \quad \int_{S_h} w_h \cdot \nabla v = 0, \text{ for } v \in S_h \text{ with support in } \mathbb{R}^2.$$

We shall show that then

$$(24) \quad \|w_h\|_{L_2(\Omega)} \leq C d^{-1} \|w_h\|_{L_2(\mathbb{R}^2)}.$$

Comment.

For our proof to work, we need that  $d$  is so large that every triangle intersects at most one of the circular boundary pieces of  $\Omega^k$ ,  $k = 2, 3, 4$ , and  $'$ . This can be arranged by taking  $C_*$  large enough. (cf. the comment after (17).)

Introduce a smooth cut-off function  $\omega(x)$ ,  $x \in \mathbb{R}^2$ ,  $0 \leq \omega \leq 1$ , such that

$$1 \text{ on } A, \quad \text{supp } \omega \subseteq A^4$$

and such that

$$(25) \quad \|w_\ell\|_{L_2(\mathbb{R}^2)} \leq C d^{-\ell}, \quad \ell = 0, 1, 2.$$

Such a function is easily constructed by change of variables in one valid for  $d=1$ .

Recall our notational convention following (14)),

$$(26) \quad \| \nabla w_h \|_{L_2(\Omega)}^2 = \| \omega \nabla w_h \|_{L_2}^2.$$

Here,

$$\begin{aligned} \| \nabla w_h \|_{L_2}^2 &= \int_{\Omega}^2 \nabla w_h \cdot \nabla w_h \\ &= \int_{\Omega}^2 \omega \nabla w_h \cdot (\omega \nabla w_h) - 2 \int_{\Omega}^2 \omega \nabla w_h \cdot (\omega \nabla w_h) F_1 + F_2. \end{aligned}$$

With  $I$  the interpolant of  $\omega^2 w_h$  ( $I$  is supported in  $\Omega^3$  for  $C_*$  large), using the discrete harmonicity of  $w_h$  we have by (22) and (25)

$$\begin{aligned} |F_1| &= \left| \int_{\Omega^3} \nabla w_h \cdot (\omega^2 w_h - I) \right| \\ &\leq C \| \nabla w_h \|_{L_2(\Omega^3)} h^{(d-2)} \| w_h \|_{L_2(\Omega^3)} + d^{-1} \| \omega \nabla w_h \|_{L_2} \end{aligned}$$

Using now the well known inverse property that

$h^{-1} \leq c_{\Omega}^{-1}$ , we have upon squaring and summing over all elements  $\tau$  intersecting  $\Omega^3$ ,

$$Ch \| \nabla w_h \|_{L_2(\Omega^3)} \leq C \| w_h \|_{L_2(\Omega^2)} \quad (C_* \text{ large enough}).$$

Further,

$$|F_2| \leq C d^{-1} \| \omega \nabla w_h \|_{L_2} \| w_h \|_{L_2(\Omega^4)}.$$

Collecting the above estimates, since  $\Omega^4 \subset \Omega^3 \subset \Omega^2$ ,

$$\begin{aligned}\|\omega^\alpha w_h\|_{L_2}^2 &\leq C d^{-2} \|w_h\|_{L_2(\Omega^3)}^2 + C d^{-1} \|w_h\|_{L_2(\Omega^2)} \|\omega^\alpha w_h\|_{L_2} \\ &\leq C d^{-2} \|w_h\|_{L_2(\Omega^2)}^2 + \frac{1}{2} \|\omega^\alpha w_h\|_{L_2}^2\end{aligned}$$

whereupon, by (26), the desired estimate (24) obtains.

We proceed in pursuit of Lemma 3. We now employ a cut-off function  $\eta(x)$  such that

$$\eta \equiv 1 \text{ on } A^2, \quad \text{supp } \eta \subseteq A',$$

and satisfying the estimate of (25). Then

$$\begin{aligned}\|\nabla v_h\|_{L_2(\Omega)} &\leq \|\nabla(P_h(\eta v))\|_{L_2} + \|\nabla(P_h(\eta v) - v_h)\|_{L_2(\Omega)} \\ &\leq \|\nabla(\eta v)\|_{L_2} + \|\nabla(P_h(\eta v) - v_h)\|_{L_2(\Omega)},\end{aligned}$$

since the projection obviously is stable in the energy (quasi) norm over  $R_h$ . Hence,

$$(27) \quad \|\nabla v_h\|_{L_2(\Omega)} \leq C \|\nabla v\|_{L_2(\Omega')} + C d^{-1} \|v\|_{L_2(\Omega')} + \|\nabla(P_h(\eta v) - v_h)\|_{L_2(\Omega)}.$$

For the third term on the right of (27), since  $\eta \equiv 1$  on  $A^2$ ,  $P_h(\eta v) - v_h$  is discrete harmonic on  $\Omega^2$ . Therefore from (24),

$$\begin{aligned}\|\nabla(P_h(\eta v) - v_h)\|_{L_2(\Omega)} &\leq C d^{-1} \|P_h(\eta v) - v_h\|_{L_2(\Omega^2)} \\ &\leq C d^{-1} \|P_h(\eta v) - \eta v\|_{L_2(\Omega^2)} + C d^{-1} \|v - v_h\|_{L_2(\Omega^2)}.\end{aligned}$$

Combining this with (27), using also the triangle inequality,

$$\begin{aligned}
 \| \nabla e \|_{L_2(\Omega)} &\leq \| \nabla v \|_{L_2(\Omega)} + \| \nabla v_h \|_{L_2(\Omega)} \leq C \| \nabla v \|_{L_2(\Omega')} + Cd^{-1} \| v \|_{L_2(\Omega')} \\
 (28) \quad &+ Cd^{-1} \| e \|_{L_2(\Omega')} \\
 &+ Cd^{-1} \| P_h(nv) - nv \|_{L_2(\Omega')}
 \end{aligned}$$

To handle the last term on the right we use a duality argument:

$$\begin{aligned}
 \| P_h(nv) - nv \|_{L_2(\Omega')} &= \sup_{\phi \in C_0^\infty(\Omega')} \int (P_h(nv) - nv) \phi . \\
 (29) \quad &\| \phi \|_{L_2} = 1
 \end{aligned}$$

For each such  $\phi$ , let  $\psi$  be the solution of

$$-\Delta \psi = \phi \quad \text{in } R, \quad \psi = 0 \quad \text{on } \partial R.$$

Then since  $P_h(nv) = 0$  on  $\partial R_h$ , Green's formula gives

$$(30) \quad \int_{R_h} (P_h(nv) - nv) \phi = \int_{R_h} \nabla (P_h(nv) - nv) \cdot \nabla \psi - \oint_{\partial R_h} nv \frac{\partial \psi}{\partial n} \equiv I_1 + I_2 .$$

For  $I_1$ : By well known properties,

$$\begin{aligned}
 |I_1| &= \left| \int_{R_h} \nabla (P_h(nv) - nv) \cdot \nabla (\psi - P_h \psi) \right| = \left| \int_{R_h} \nabla (nv) \cdot \nabla (\psi - P_h \psi) \right| \\
 (31) \quad &\leq C \| \nabla (nv) \|_{L_2} \| \psi \|_{H^2(R)} \\
 &\leq Ch (\| \nabla v \|_{L_2(\Omega')} + d^{-1} \| v \|_{L_2(\Omega')}) .
 \end{aligned}$$

For  $I_2$ : Note that the term enters only if  $\partial R_h$  intersects  $\text{supp}(n)$ .

We have

$$|I_2| \leq \|\eta v\|_{L_2(\partial R_h)} \|\nabla \psi\|_{L_2(\partial R_h)} .$$

Since  $\partial R_h$  is uniformly Lipschitz (easily checked),

$$\begin{aligned} \|\eta v\|_{L_2(\partial R_h)} &\leq C(\|\eta v\|_{L_2} \|\eta v\|_{H^1})^{1/2} \\ (32) \quad &\leq C(d^{-1/2} \|v\|_{L_2(\Omega')} + d^{1/2} \|\nabla v\|_{L_2(\Omega')}) . \end{aligned}$$

Further,

$$(33) \quad \|\nabla \psi\|_{L_2(\partial R_h)} \leq C(\|\nabla \psi\|_{L_2} \|\psi\|_{H^2})^{1/2} ,$$

where

$$(34) \quad \|\psi\|_{H^2} \leq C.$$

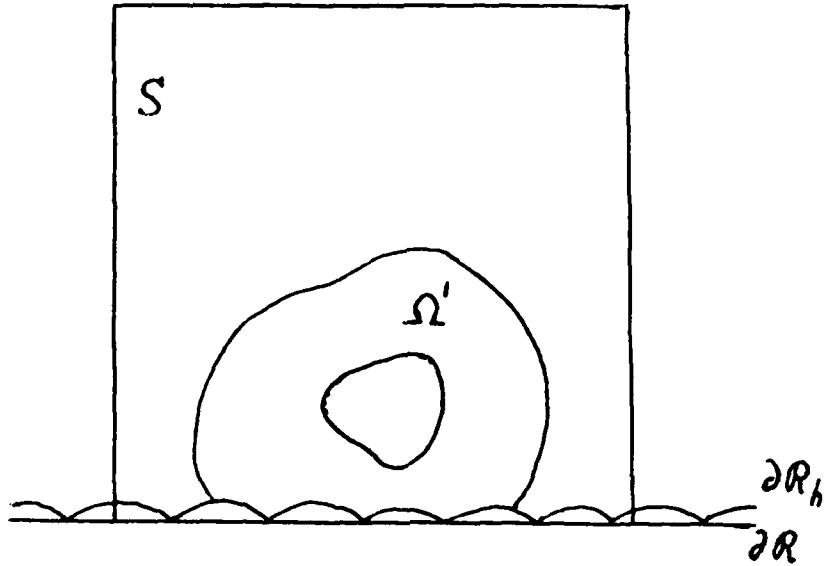
For the other factor in (33), we have, taking the norm over the whole of  $R$ ,

$$(35) \quad \|\nabla \psi\|_{L_2(R)}^2 = \int_{\Omega'} \psi \phi \leq \|\psi\|_{L_2(\Omega')} .$$

We next want to show that

$$(36) \quad \|\psi\|_{L_2(\Omega')} \leq Cd \|\nabla \psi\|_{L_2(R)} .$$

Since  $\partial R_h$  and  $\text{supp}(\eta)$  intersect we have the following picture, after straightening the boundary  $\partial R$  in a neighborhood, assumed greater than  $d$ .



Here the domain  $S$  is a square of side length  $\leq Cd$  that contains  $\Omega'$ . By Poincaré's inequality then, since  $\psi$  vanishes on  $\partial R$ , we obtain  $\|\psi\|_{L_2(S)} \leq Cd \|\nabla \psi\|_{L_2(S)}$  and thus (36).

There are only finitely many  $d$  ( $=d_j$ ) not covered by the above and for these we obtain (36) from Poincaré's inequality over the whole of  $R$  (possibly increasing the constant).

From (35) and (36),  $\|\nabla \psi\|_{L_2} \leq Cd$ , and reporting this and (34) into (33), we get

$$\|\nabla \psi\|_{L_2(\partial R_h)} \leq C d^{1/2}.$$

Using this and (32),

$$|I_2| \leq C(d\|\nabla v\|_{L_2(\Omega')} + \|v\|_{L_2(\Omega')}).$$

By this and (31) in (30) and (29), since  $d \geq ch$ ,

$$\|P_h(v)-v\|_{L_2(\Omega')} \leq C(d\|\nabla v\|_{L_2(\Omega')} + \|v\|_{L_2(\Omega')}).$$

whereupon inserting into (28), we obtain

$$\|\nabla(v-v_h)\|_{L_2(\Omega)} \leq C(\|\nabla v\|_{L_2(\Omega')} + d^{-1}\|v\|_{L_2(\Omega')} + d^{-1}\|v-v_h\|_{L_2(\Omega')}).$$

This estimate depends only on the fact that  $v_h = P_h v$ . Therefore we may write  $v-v_h = v-x-P_h(v-x)$  for any  $x$  in  $S_h$  and obtain

$$\begin{aligned} \|\nabla(v-v_h)\|_{L_2(\Omega)} &\leq C \min_{x \in S_h} (\|\nabla(v-x)\|_{L_2(\Omega')} + d^{-1}\|v-x\|_{L_2(\Omega')}) \\ &\quad + Cd^{-1}\|v-v_h\|_{L_2(\Omega')} \end{aligned}$$

This proves the lemma, and ends the lecture.

Lars B. Wahlbin, Cornell University.

Third Lecture: A brief survey of parabolic smoothing and how it affects a numerical solution: finite differences and finite elements.

Finite Differences.

As a suitable model problem we take the pure Cauchy problem for the heat equation in one space variable, i.e., the problem of finding  $u = u(t, x)$  such that

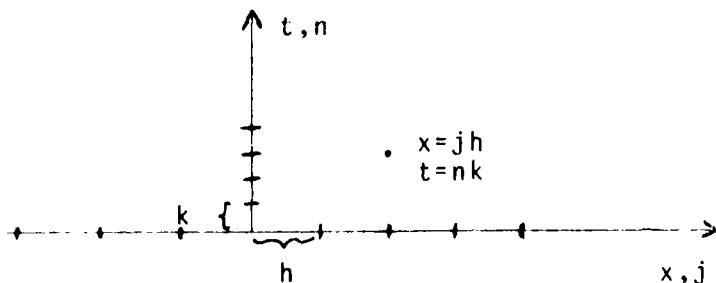
$$(1) \quad \begin{cases} u_t = u_{xx}, & -\infty < x < \infty, t > 0, \\ u(0, x) = v(x). \end{cases}$$

Here  $v$  is a given function.

Define the solution operator  $E(t)$  by

$$(2) \quad u(t, x) = E(t)v(x) = \frac{1}{(4\pi t)^{1/2}} \int e^{-(x-y)^2/4t} v(y) dy.$$

For numerical solution of (1) we choose a regular mesh. With  $h$  a spatial steplength and  $k$  a temporal one we introduce notation according to the following figure.



We demand that, as  $k$  and  $h$  vary, they obey

$$\lambda = k/h^2 \text{ is fixed.}$$

As a model example, consider the forward time, centered space

Lecture in the Department of Mathematics at University of Maryland, College Park, during their Special Year in Numerical Analysis, 1980-81.

approximation (FTCS). With  $u_j^n$  the approximate solution,

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2},$$

or

$$u_j^{n+1} = \lambda u_{j+1}^n + (1-2\lambda)u_j^n + \lambda u_{j-1}^n = (E_k u^n)_j.$$

For the moment, set  $u_j^0 = v_j = v(jh)$ ; we shall return to the question of how to choose the discrete initial data. Then, corresponding to (2), we have the discrete solution operator

$$u_j^n = (E_k^n v)_j.$$

We shall describe some results for this set-up. Sharp theorems for the rate of convergence in terms of smoothness of initial data  $v$  are obtained by use of the Fourier transform, and of Besov spaces. We have

$$\widehat{E(t)v}(\xi) = e^{-t\xi^2} \hat{v}(\xi),$$

$$\widehat{E_k^n v}(\xi) = a(h\xi)^n \hat{v}(\xi),$$

where, in the FTCS case,

$$a(\theta) = 1-2\lambda+2\lambda \cos(\theta).$$

In general,  $a(\theta)$  is  $2\pi$ -periodic, and we demand parabolicity in the sense of John, [1], i.e., that there exists a positive constant  $c$  such that

$$|a(\theta)| \leq e^{-c\theta^2}, \text{ for } |\theta| \leq \pi.$$

In the FTCS approximation, this is the case if  $\lambda < 1/2$ ; it is well known that the approximation is useless for  $\lambda > 1/2$ .

We also introduce the order of accuracy,  $\mu$ , of the approximation via

$$|a(\theta) - e^{-\theta}|^2 = O(\theta^{2+}) \quad \text{as } |\theta| \rightarrow 0.$$

In the FICCS scheme,  $\nu = 2$ .

We next briefly describe the Besov spaces  $B_p^s (= B_p^{s,\infty}(R^1))$  to the extent that the audience can appreciate that they might come in handy in a Fourier based investigation. Loosely speaking, a function in  $B_p^s$  has (almost)  $s$  derivatives in  $L_p$ . A convenient characterisation for the present purposes is the following; cf. [2] and references there for details.

Let

$$v = \phi_0(z) + \phi_1(z) + \dots + \phi_i(z) + \dots$$

where

$\phi_0$  is a smooth characteristic function of the interval  $[-1,1]$ ,

$\phi_i$ ,  $i \geq 1$ , is a smooth characteristic function of  
 $\{2^{i-1} \leq |z| \leq 2^{i+1}\}$ .

Then

$$\|v\|_{B_p^s} = \sup_{i \geq 0} 2^{si} \|F^{-1}(\phi_i \hat{v})\|_{L_p}.$$

For  $p = \infty$  we have,

$s$  non-integral,  $B_\infty^s = \text{Lip}(s)$ , the Hölder class,

$s$  integral,  $B_\infty^s = \text{Zyg}(s)$ , the Zygmund class.

The following result characterises the rate of convergence in terms of the smoothness of initial data  $v$ . It takes into account all possible translations of the spatial mesh, and the error at all time-levels  $n_k$ .

THEOREM 1. [3].

For  $0 < s \leq \mu$ ,

$$\|E_k^n v - E(nk)v\|_{L_\infty} \leq Ch^s \|v\|_{B_\infty^s}, \quad n = 1, 2, \dots$$

Conversely, if for a fixed  $t_1 > 0$ ,

$$\sup_{0 < nk < t_1} \|E_k^n v - E(nk)v\|_{L_\infty} \leq Ch^s \text{ as } h, k \rightarrow 0,$$

then  $v \in B_\infty^s$ .

This result should be compared with classical results for trigonometric approximation, cf. [3] and references there, and results for spline approximation, [4].

A generalization of the first part of Theorem 1 to more complicated problems can be found in [5].

Example.

Let  $\sigma$  be a small positive number and  $v_\sigma$  the function

$$(3) \quad v_\sigma(x) = \begin{cases} 0, & x < 0, \\ x^\sigma \omega(x), & x \geq 0, \end{cases}$$

where  $\omega(x)$  is smooth function with compact support on  $R^1$ , with  $\omega \equiv 1$  in a neighborhood of the origin. Then, see e.g. [2, Prop 2.4.2],  $v_\sigma \in B_\infty^0$  but  $v_\sigma \notin B_\infty^s$  for  $s > \sigma$ . Here we have by a simple calculation,

$$E(k)v_\sigma(0) = c_\sigma \lambda^{\sigma/2} h^\sigma + O(e^{-1/(10\lambda h^2)})$$

and, for the FTCS scheme,

$$E_k^1 v_\sigma(0) = \lambda h^\sigma.$$

Rightfully then, Theorem 1 predicts only  $h^\sigma$  convergence close

to initial time (unless the mesh parameter  $\lambda$  is luckily chosen; then shift the mesh).

However, let us now take into account the smoothing property of (1), i.e., the fact that even if  $v$  is rough,  $E(t)v$  is smooth for  $t > 0$ . Will the numerical solution take advantage of that? Let us therefore study convergence at a fixed positive time  $t_0 = nk$ . Our first theorem may come as a disappointment.

#### THEOREM 2. [6].

Let  $0 < s \leq 1$ , and let  $t_0 = nk > 0$  be fixed. There exists a function  $v \in B_\infty^s$  such that

$$\limsup_{\substack{h,k \rightarrow 0 \\ nk=t_0}} h^{-s} \|E_k^n v - E(nk)v\|_{L_\infty} > 0.$$

A specific such function was exhibited in [7]; it is, essentially, a lacunary Fourier series. It is not particularly likely to come up in many applications. E.g., it has the property, for  $s$  a non-zero integer,  $d^{s-1} v/dx^{s-1}$  is continuous, but  $d^s v/dx^s$  is non-existent a.e.

An attempt at an inverse theorem for convergence at a fixed time leads to the following.

#### THEOREM 3. [7].

Let  $1 < s \leq \mu$ , and let  $t_0 = nk > 0$  be fixed. Then if  $\|E_k^n v - E(nk)v\|_{L_\infty} \leq Ch^s$ , as  $h,k \rightarrow 0$ , we have  $v \in B_\infty^{s-1}$ .

A positive counterpart to the above is as follows.

THEOREM 4. [7].

Let  $1 < s \leq \mu$ . Then

$$\| E_k^n v - E(nk)v \|_{L_\infty} \leq Ch^s (nk)^{-1/2} \| v \|_{B_1^s}.$$

Example.

Consider the function  $v_\sigma$  of (3). It belongs to  $B_1^{1+\sigma}$  and to  $B_\infty^\sigma$ , but no better. Theorem 4 predicts  $O(h^{1+\sigma})$  convergence ( $0 < \sigma \leq \mu - 1$ ) for positive time, and Theorem 3 tells us that this is sharp.

Hence, the combination of Theorem 3 and Theorem 4 is sharp for this type of isolated roughness.

The fact that the numerical approximation takes advantage of the smoothing property was noted in [8].

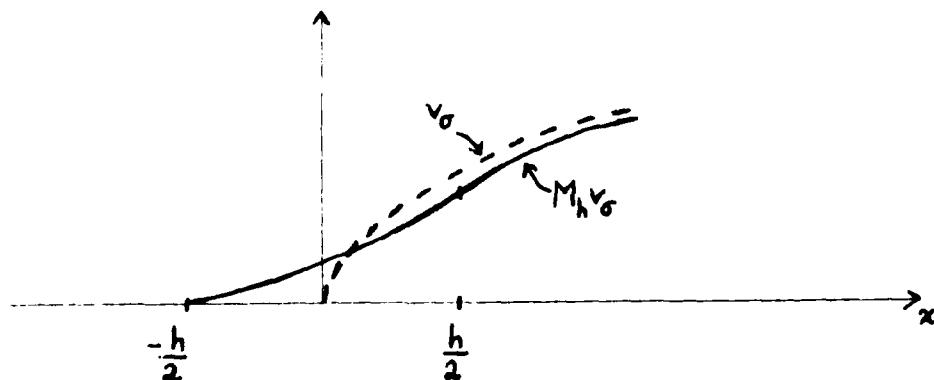
Still better rate of convergence for positive time than predicted by Theorem 4 can be accomplished if one employs a preliminary smoothing of initial data, [7], [9].

Example.

We define a mean-value operator  $M_h$ ,

$$M_h v(x) = \frac{1}{h} \int_{-h/2}^{h/2} v(x-y) dy = F^{-1}\left(\frac{2 \sin(h\xi/2)}{h\xi} \hat{v}(\xi)\right)(x).$$

We have, e.g., the following picture for  $v_\sigma$  of (3).



Here

$$(4) \quad M_h v_\sigma(x) = \begin{cases} 0, & x < -h/2 \\ \frac{1}{h(\sigma+1)} (x + h/2)^{\sigma+1}, & -h/2 \leq x \leq h/2 \\ \frac{1}{h(\sigma+1)} ((x + h/2)^{\sigma+1} - (x - h/2)^{\sigma+1}), & h/2 < x, \end{cases}$$

as long as  $\omega = 1$ .

THEOREM 5. [7].

For  $1 < s \leq n$ ,  $0 < t_0 < t_1$  fixed,

$$\| E_k^n M_h v - E(nk)v \|_{L_\infty} \leq Ch^2 \| v \|_{B_1^s}, \text{ for } 0 < t_0 \leq nk \leq t_1.$$

Applying this result to  $v_\sigma$ , while taking the difference scheme directly on  $v_\sigma$  gives a rate of  $O(h^{1+\sigma})$ , using it on  $M_h v_\sigma$  results in an  $O(h^2)$  rate. Thus, for  $\sigma$  small, we gain almost a full order.

The simple mean-value operator considered fits nicely with the FTCS approximation; the inherent second order in the scheme is restored as long as initial data has somewhat more than one derivative, the derivative being measured in the weakest possible  $L_p$ -class, viz., in  $L_1$ . Similar results hold for higher order schemes by use of higher order smoothing operators.

Note that if it is possible to decompose initial data  $v$  as  $v = v_{\text{smooth}} + v_{\text{rough}}$ , then it is sufficient to apply the smoothing operator to  $v_{\text{rough}}$ . If one is lucky one might choose  $v_{\text{rough}}$  so that the smoothing operator on it is easily evaluated, cf. (4), where one could use a smooth spline cut-off for  $m$  in (3).

Theorem 5 holds also for more complicated problems, [7].

We conclude our survey of results in the finite difference theory by giving a very rough indication of why a smoothing operator works:

Split  $v$  as  $\hat{v} = \hat{v}_{\text{low}} + \hat{v}_{\text{high}}$  into "low" and "high" frequency components. For rough  $v$ ,  $\hat{v}_{\text{high}}$  is not "small". With  $t_0 = nk = \lambda nh^2$ ,

$$(E_k^n v - E(nk)v)^\wedge(\xi) = (a(h\xi)^n - e^{-\lambda n(h\xi)^2})\hat{v}_{\text{low}}$$

$$- e^{-t_0 \xi^2} \hat{v}_{\text{high}}$$

$$+ a(h\xi)^n \hat{v}_{\text{high}} = I_1 + I_2 + I_3.$$

Here,  $I_1$  is "small" by the order of accuracy condition. Clearly,  $I_2$  is "small" for  $t_0$  fixed positive; this reflects the smoothing property of (1). However, since  $a(h\xi) = 1$  for  $h\xi = 2i\pi$ , i integer,  $I_3$  will not be small.

Considering next what happens when applying  $M_h$ , in the same sloppy manner we have

$$(E_k^n M_h v - E(nk)v)^\wedge(\xi)$$

$$= a(h\xi)^n [\frac{\sin(h\xi/2)}{h\xi/2} - 1] \hat{v}_{\text{low}}$$

$$+ (a(h\xi)^n - e^{-\lambda n(h\xi)^2})\hat{v}_{\text{low}}$$

$$- e^{-t_0 \xi^2} \hat{v}_{\text{high}}$$

$$+ a(h\xi)^n \frac{\sin(h\xi/2)}{h\xi/2} \hat{v}_{\text{high}} = J_1 + J_2 + J_3 + J_4.$$

Here,  $J_1$  is "small" since  $\frac{\sin \theta}{\theta} \sim 1$  for  $\theta$  small.  $J_2$  and  $J_3$  are "small" for the same reason  $I_1$  and  $I_2$  were. Finally, for  $J_4$ , the role of the smoothing operator in damping high frequency components is easily discerned.

#### Finite Elements.

As a suitable model problem, this time we consider the mixed

initial boundary value problem

$$(5) \quad \begin{cases} u_t = u_{xx}, & 0 < x < 1, \quad t > 0, \\ u(t,0) = u(t,1) = 0, \\ u(0,x) = v(x). \end{cases}$$

Let  $u(t) = E(t)v$ .

For the numerical set-up, let  $N$  be a sequence of integers,  $h = 1/N$ ,  $I_j = [jh, (j+1)h]$ ,  $j=0, \dots, N-1$ ; let  $r$  and  $s$  be integers with  $r \geq 2$ ,  $0 \leq s \leq r-2$  and set

$$S_h = \{x \in C^s[0,1], x(0) = x(1) = 0, x|_{I_j} \text{ polynomial of degree } r-1\}.$$

For simplicity we only consider a semi-discrete approximation to (5). We seek  $u_h(t) \in \overset{\circ}{S}_h$  such that

$$\begin{cases} ((u_h)_t, x) + ((u_h)_x, x_x) = 0, & \text{for all } x \in \overset{\circ}{S}_h, \\ u_h(0) = v_h \text{ given in } \overset{\circ}{S}_h. \end{cases}$$

Here  $(f,g) = \int_0^1 fg$ . Set  $E_h(t)v_h = u_h(t)$ .

In this situation, let us not review results of a general nature (cf. [10], [11], [12]) but move directly to the problem of how the choice of  $v_h$  influences the "smoothing advantage". Furthermore, the precise results we state shall only be quoted for (maximal)  $O(h^r)$  convergence.

The following very loose indication can be given. Let  $\phi_j(x) = \sqrt{2} \sin(\pi jx)$ ,  $\lambda_j = \pi^2 j^2$ ,  $j = 1, 2, \dots$  be the eigenfunctions and eigenvalues of the operator  $-d^2/dx^2$  on  $[0,1]$  with zero endpoint conditions; let  $\phi_j^h$  and  $\lambda_j^h$ ,  $j = 1, \dots, \approx N$ , be their discrete counterparts. They are "close" for "low"  $j$ , but unrelated for

"high"  $j$ , although for  $j$  large,  $\lambda_j^h$  is large. Now,  
 $E(t)v = \sum_j k_j e^{-\lambda_j t} \phi_j$ ,  $k_j = (v, \phi_j)$ ,

and

$$E_h(t)v = \sum_j k_j^h e^{-\lambda_j^h t} \phi_j^h, k_j^h = (v_h, \phi_j^h).$$

Therefore,

$$\begin{aligned} E_h(t)v_h - E(t)v &= \sum_{j \text{ low}} (k_j^h - k_j) e^{-\lambda_j^h t} \phi_j^h \\ &+ \sum_{j \text{ low}} k_j (e^{-\lambda_j^h t} \phi_j^h - e^{-\lambda_j t} \phi_j) + \sum_{j \text{ high}} k_j^h e^{-\lambda_j^h t} \phi_j^h \\ &- \sum_{j \text{ high}} k_j e^{-\lambda_j t} \phi_j = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Here  $I_2$  is "small" because  $\lambda_j^h \sim \lambda_j$  and  $\phi_j^h \sim \phi_j$  for low  $j$ . For fixed positive  $t$ ,  $I_3$  and  $I_4$  are "small" since  $\lambda_j$  and  $\lambda_j^h$  are large.

To have  $I_1$  small, we need that  $k_j^h - k_j$  is small for low  $j$ .

We write

$$k_j^h - k_j = (v_h - v, \phi_j^h) - (v, \phi_j^h - \phi_j), \text{ low } j,$$

where the second member on the right is "small". Thus, loosely speaking, we need

$$(6) \quad (v_h - v, \phi_j^h) \text{ small, for low } j.$$

If we take  $v_h$  to be the interpolant of  $v$ , then (6) would hold provided  $v$  was smooth; for rough  $v$ , systematic sign errors of large magnitude could occur, destroying (6).

There is, however, the obvious choice of setting  $v_h$  to be the  $L_2$  projection of  $v$ .

The above can be made exact. Let  $P_h v$  denote the  $L_2$ -projection, defined by  $(P_h v, x) = (v, x)$  for all  $x \in \overset{\circ}{S}_h$ .

THEOREM 6. [13].

Let  $t_0 > 0$  and  $v_h = P_h v$ . Then

$$\|E_h(t)P_h v - E(t)v\|_{L_\infty} \leq Ch^r \|v\|_{L_1}, \quad 0 < t_0 \leq t.$$

A similar result holds in the presence of a suitable time discretization, [11], [13].

The smoothing property in parabolic finite element equations was investigated in [14] and [15].

In the case of smoothest splines ( $s=r-2$ ) there is a connection between the Galerkin method and certain "finite difference" operators, see [16]. Then taking  $v_h = P_h v$  corresponds to a smoothing operator in the finite difference theory.

In general, numerical integration is needed to evaluate  $P_h v$ . Let, cf. [17] for details,  $\tilde{P}_h v$  denote the approximate  $L_2$ -projection given by applying an integration method which is exact, on each subinterval, for polynomials of degree  $2r-2$ .

THEOREM 7. [17].

Let  $t_0 > 0$  and  $v_h = \tilde{P}_h v$ . Then

$$\|E_h(t)\tilde{P}_h v - E(t)v\|_{L_\infty}$$

$$\leq Ch^r \{ \|v\|_{L_\infty} + \sum_{j=0}^{N-1} \int_{I_j} \min(x, 1-x) |D^r(x)| dx \}$$

$$+ \sum_{j=0}^{N-1} \int_{I_j} |D^{r-1}v(x)| dx \}, \quad \text{for } 0 < t_0 \leq t.$$

A similar result holds if  $v_h$  is the interpolant of  $v$ ; the

result also carries over to time-discretizations, [17].

Applying Theorem 7 to functions behaving like  $(x-x_0)_+^\sigma$  locally, we have for  $\sigma$  non-integral and  $x_0$  interior, that  $\sigma > r-1$  is needed for  $O(h^r)$  convergence. This corresponds to Theorem 4 in the finite difference situation. For  $x_0 = 0$  or 1, or for  $\sigma$  integer and  $x_0$  meshpoint we have more advantageous estimates in the finite element situation; this is because the analysis does not have to include all possible mesh-shifts.

As in the case of preliminary smoothing in the finite difference setting, if  $v$  can be split as  $v = v_{\text{smooth}} + v_{\text{rough}}$ , it would suffice to evaluate  $P_h v_{\text{rough}}$  exactly.

We conclude this lecture with the following five brief remarks.

i) Take e.g.  $v$  to be the step function

$$v(x) = \begin{cases} 0, & 0 \leq x \leq 1/4, \\ 1, & 1/4 \leq x \leq 3/4, \\ 0, & 3/4 \leq x \leq 1. \end{cases}$$

The  $L_2$  projection has then an oscillatory error, which gets heavily damped in the approximate solution, which in this respect behaves like the true solution.

ii) Connected with i) is the fact that

$$\|v - P_h v\|_{H^{-r}} \leq C h^r \|v\|_{L_2}.$$

iii) A smoothing property in the Navier-Stokes equations has been put to a somewhat similar use in [18].

iv) For the Euler-Poisson-Darboux equation,

$$u_{tt} + \frac{K}{t} u_t = \Delta u$$

there is a smoothing property, which depends on the size of  $K$ . The

finite element solution takes advantage of this, [19].

v) The influence of time-discretizations on the parabolic smoothing property in finite elements has been thoroughly investigated in [20], [21]. Certain surprises are in store for high order time-discretizations when the equation has time dependent coefficients.

\*

#### References.

1. F. John, On integration of parabolic equations by difference methods. I. Linear and quasi-linear equations for the infinite interval, Comm. Pure Appl. Math. 5, 1952, 155-211.
2. P. Brenner, V. Thomée and L.B. Wahlbin, Besov spaces and Applications to Difference Methods for Initial Value Problems, Lecture Notes in Mathematics 434, Springer, New York, 1975.
3. J. Löfström, Besov spaces in the theory of approximation, Ann. Mat. Pura Appl., 85, 1970, 93-184.
4. O. Widlund, On best error bounds for approximation by piecewise polynomial functions, Numer. Math. 27, 1977, 327-338.
5. O. Widlund, On the rate of convergence for parabolic difference schemes. II, Comm. Pure Appl. Math. 23, 1970, 79-96.
6. G.W. Hedstrom, The rate of convergence of some difference schemes, J. SIAM Numer. Anal. 5, 1968, 363-406.
7. V. Thomée and L. Wahlbin, Convergence rates of parabolic difference schemes for non-smooth data, Math. Comp. 28, 1974, 1-13.
8. M.L. Juncosa and D.M. Young, On the Crank-Nicolson procedure for solving partial differential equations, Proc. Cambridge Philos. Soc. 53, 1957, 448-461.

9. H.O. Kreiss, V. Thomée and O. Widlund, Smoothing of initial data and rates of convergence for parabolic difference equations, Comm. Pure Appl. Math. 23, 1970, 241-259.
10. J.H. Bramble, A.H. Schatz, V. Thomée and L.B. Wahlbin, Some convergence estimates for semi-discrete Galerkin type approximations for parabolic equations, SIAM J. Numer. Anal. 14, 1977, 218-241.
11. G.A. Baker, J.H. Bramble and V. Thomée, Single step Galerkin approximations for parabolic problems, Math. Comp. 31, 1977, 818-847.
12. A.H. Schatz, V. Thomée and L.B. Wahlbin, Maximum norm stability and error estimates in parabolic finite element equations, Comm. Pure Appl. Math. 33, 1980, 265-304.
13. V. Thomée, Some convergence results for Galerkin methods for parabolic boundary value problems, Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, 55-88.
14. H.-P. Helfrich, Fehlerabschätzungen für das Galerkinverfahren zur Lösung von Evolutionsgleichungen, Manuscripta Math. 13, 1974, 219-235.
15. H. Fujita and A. Mizutani, On the finite element method for parabolic equations, I: Approximation of holomorphic semigroups, J. Math. Soc. Japan, 28, 1976, 749-771.
16. V. Thomée, Spline approximation and difference schemes for the heat equation, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A.K. Aziz, ed., Academic Press, New York, 1972, 711-746.
17. L.B. Wahlbin, A remark on parabolic smoothing and the finite element method, SIAM J. Numer. Anal. 17, 1980, 33-38.

18. J.G. Heywood and R. Rannacher, Finite element approximation of the nonstationary Navier-Stokes problem, Part 3, to appear.
19. A.M. Genis, On finite element methods for the Euler-Poisson-Darboux equation, Thesis, Cornell University, 1980.
20. P. Sammon, Approximations for parabolic equations with time-dependent coefficients, Thesis, Cornell University, 1978.
21. P. Sammon, Fully discrete approximation methods for parabolic problems with nonsmooth initial data, to appear.

I. Asymptotic Convergence of  
Boundary Element Methods  
and  
II. Integral Equation Methods for  
Mixed Boundary Value Problems

by

Wolfgang L. Wendland  
Technische Hochschule Darmstadt,  
Germany, Federal Republic

Lecture Notes on two lectures given at the University of Maryland in March 1981 and also at the University of Delaware in May 1981.

Acknowledgements

The work in these lectures has been supported jointly by the University of Maryland, the Applied Mathematics Institute and the Department of Mathematical Sciences of the University of Delaware and the German Research Foundation DFG (No. We 659/4-1). The author also wants to thank J. F. Harris for the typing of these notes.

## I. Asymptotic Convergence of Boundary Element Methods

This lecture gives a survey on the asymptotic error analysis of boundary integral equations, in particular on joint work by M. Costabel G. C. Hsiao, P. Kopp, E. Stephan and W. L. Wendland [25, 34, 38, 39, 41, 42, 43, 72, 79, 80].

### Introduction

Nowadays the most popular numerical methods for solving elliptic boundary value problems are finite differences [22], finite elements [11, 21] and, more recently, boundary integral methods. The latter are numerical methods for solving integral equations (or their generalizations) on the boundary of the given domain.

The conversion of elliptic boundary value problems into corresponding integral equations for the investigation of existence goes far back in history. For computational purposes, however, boundary integral equations of various types became more fashionable only recently. (See e.g. [44] and the proceedings [16, 17] and [68].) The numerical discretizations are mostly based on collocation methods whereas mathematically Galerkin's procedure and corresponding variational formulations provide a further developed error analysis.

Here we shall develop a numerical implementation of Galerkin's procedure. The resulting scheme not only provides high accuracy as Galerkin's method but also is simple to be adapted to modern computing machines. We shall term this method as the Galerkin collocation method [79].

It applies to a very wide class of integral equations on the boundary manifold  $\Gamma$  as to integral equations of the second and the first kind, to singular integral equations with Cauchy kernels on curves and Giraud kernels on surfaces, i.e. Calderon Zygmund operators [19] and also some integrodifferential equations with finite part principal value operators.

The method generalizes the Galerkin collocation in [38] that has been developed for Fredholm integral equations of the first kind with the logarithmic kernel as the principal part.

The effectiveness of the method rests on the asymptotic convergence properties of Galerkin's method. For finite element methods in the domain and for finite differences it is well known that strong ellipticity implies the asymptotic convergence. But for the boundary integral methods the strong ellipticity of the corresponding pseudodifferential operators seemed not to have received the proper attention yet.

Here we shall focusing towards (i) strong ellipticity, (ii) a priori estimates for the integral equations, and for two-dimensional problems towards (iii) convolution operators as the principal parts and (iv) smoothness of the remaining kernels.

(i) Strong ellipticity:

Since Michlin's fundamental work [51] and the constructive proof of the Lax-Milgram theorem by Hildebrandt and Wienholtz [35] it is well known that the Gårding inequality, i.e. strong ellipticity implies asymptotic convergence of Galerkin's method in the energy norm. This in turn gives optimal convergence rates

in the corresponding Sobolev spaces. Using  $L_2$  or the Sobolev space norms which are equivalent to the energy norm it turns out that the strong ellipticity is even necessary for the convergence of all Galerkin procedures due to Vainikko [77]. As for the variational methods [21], the use of regular finite element functions yields optimal order of convergence when the error is measured in the energy norm or in Sobolev space norms of higher order. (See also Stephan and Wendland [72].) Here we consider equations which are strongly elliptic with ellipticity corresponding to Agmon, Douglis and Nirenberg (see [36] p. 268) but also with pseudodifferential operators of arbitrary real orders.

It should be pointed out that strong ellipticity is a rather strong condition. Often serve more specific weaker properties of the problem for satisfying the Babuška-Brezzi conditions [10].

(ii) a priori estimates:

If the integral equations are interpreted as strongly elliptic pseudodifferential equations [46, 67, 76] then they provide a priori estimates in the whole scale of Sobolev spaces in addition to the Gårding inequality. This allows to generalize the Aubin-Nitsche Lemma [57] from differential equations to the general class of strongly elliptic pseudodifferential equations as done by Hsiao and Wendland in [42]. Nitsche's trick proves super approximation i.e. optimal order of convergence even if the error is measured in Sobolev space norms of order less than the energy norm. This super approximation implies high convergence rates for the approximate potentials in compact subdomains away

from the boundary manifold where the integral equation was solved approximately. This indeed was often observed in numerical computations.

(iii) Convolution kernels as principal part:

In any case, the principal part of a pseudodifferential operator has convolutional character [67, 76]. But if it can be depicted as a simple convolution in one variable, i.e. for two-dimensional boundary value problems, then the Galerkin weights of the principal part associated with finite element functions on a regular grid form a Toeplitz matrix whose elements are given by a vector. This vector can eventually be expressed by two vectors which can be evaluated exactly up to the desired accuracy once for all independent of the boundary manifold "as well as of the meshsize" for any fixed type of finite elements. It should be pointed out that the accuracy of the numerical results depends significantly on how to compute the approximate principal part.

(iv) Smooth remaining kernels:

If the remainder of the integral operator subject to the convolutional principal part has smooth kernel then the corresponding Galerkin weights can be treated numerically by suitable quadrature formulas depending on the particular finite elements to be used and the consistency needed. This leads to simple (modified) collocation formulas.

In this way, the computation of the coefficient matrix of the finite dimensional algebraic system can be done in a most efficient and simple manner. On the other hand, the solvability of the corresponding algebraic systems as well as the asymptotic convergence of the approximate solutions are assured by the strong ellipticity of the integral equations.

If the consistency is of sufficiently high order then the asymptotic convergence and even the superapproximation remain valid for the fully discretised Galerkin collocation scheme as well.

Our replacement of the smooth part of the kernel is very much related to spline collocations of smooth kernels in Fredholm integral equations of the second kind due to Arthur [6], Chandler [20], Prenter [60] and Richter [65]. But here we are interested in an efficient approximation of the Galerkin weights rather than of the kernel due to the much wider class of equations.

Although all properties (i) - (iv) seem to restrict us to rather specific integral equations it turns out that almost all the integral equations of applications provide all these properties. In particular, the systems of integral equations of stationary and time harmonic problems of elastomechanics, thermoelasticity, of flows (viscous and inviscid) and of electromagnetics form strongly elliptic pseudodifferential equations. Several examples are listed in [79].

In Section 5 we present some numerical results from [39]. These experiments show the dependence of the accuracy on the meshwidth, i.e. the number of grid points and the smoothness of the finite elements used. In particular it can be seen that the doubling of the number of grid points yields an improvement of

one decimal digit. On the other hand, the following table shows us (for Example 5.2) that the improvement of 3 digits due to the transition from piecewise constant trial functions to continuously differentiable piecewise quadratic trial functions requires only 10% more time whereas the doubling of the number of grid points requires twice the computing time.

Table: CPU-times for examples 5.2 (both ellipses)

m	20 grid points	40 grid points
0	4.60 sec.	9.93 sec.
1	4.71 sec.	10.48 sec.
2	6.76 sec.	10.55 sec.

This comparison shows that for smooth data the use of higher order elements is more efficient than a mesh refinement.

A recent result by Prössdorf and Schmidt [62] indicates that strong ellipticity is even necessary for the convergence of Galerkin's method (with piecewise linear functions) in case of one-dimensional singular equations on a closed curve. That would mean that the projection methods of Gohberg and Feldman [32], Prössdorf [61] and Silbermann [63] with classical Fourier series converge for a wider class than our strongly elliptic equations. If one still insists on the use of finite element approximations for elliptic but not necessarily strongly elliptic equations then one has to use the least squares method [53, 72].

Similarly to differential equations, which have been treated by Bramble and Schatz [15] one again finds convergence of optimal order and super approximation [72] (for first order elliptic boundary value problems see [81] Chap. 8).

In the second lecture we shall extend our method to mixed boundary value problems where the singularities of the solution require extra care and specific approximations.

In higher dimensions, i.e. for boundary manifolds  $\Gamma \subset \mathbb{R}^n$  of dimensions  $n-1 \geq 2$  the triangulation of the manifold creates additional difficulties and additional approximations which have been studied by Nedelec [55] for a special integral equation. This was extended in [31]. In these higher dimensional cases the Toeplitz matrix of the convolutional principal part can only be defined in the above mentioned economical manner if one uses tensor product finite element functions on cubic grids in the local parametrizations. (See Michlin [52] II § 9.) All this is still to be done in detail.

Although most numerical implementations of boundary integral methods are done with the standard collocation yet there are known only few results on its asymptotic convergence except in the case of Fredholm integral equations of the second kind. Here we refer to the extensive bibliography by Ben Noble [59], the surveys by K. Atkinson [7], C. Baker [12] and results on super convergence [20, 34, 65, 66].

For our more general equations there are only preliminary results available for the special case of the Fredholm integral equation of the first kind with logarithmic kernel on the closed

boundary curve  $\Gamma$  [1, 2, 4, 78] whereas for a singular integral equation with Cauchy's kernel S. Prössdorf and G. Schmidt have proved recently that the collocation with piecewise linear functions converges if and only if the singular integral equation is strongly elliptic [62].

A more rigorous asymptotic analysis for the class of strongly elliptic equations is yet to be done. The numerical computations with boundary integral equations show super convergence where the solution is smooth. This indicates that local convergence properties also hold for the boundary integral equations.

Other open questions are uniform convergence properties and the analysis of mesh refinements and non-uniform grids on the boundary.

Since the boundary integral method is in concurrence with the well established finite element methods in the whole domain, let us make some remarks on the computational complexity for two- and three-dimensional problems. To this end let  $N$  denote the number of grid points on the simple closed boundary curve  $\Gamma$  of an interior domain  $\Omega \subset \mathbb{R}^2$  and  $N^2$  the number of grid points on the boundary surface  $\Gamma$  if  $\Omega \subset \mathbb{R}^3$ .

Then one has the following relations between complexity and  $N$  in terms of orders of  $N$ . (This comparison arose from a discussion with Professor Dr. I. Babuška, University of Maryland and Professor D. J. A. George, University of Waterloo.)

	$\Omega \subset \mathbb{R}^2$ Finite Elements in $\Omega$	$\Omega \subset \mathbb{R}^3$ Boundary integral method on $\Gamma$
Number of grid points	$N^2$	$N^3$
Stiffness matrix, computation and storage	sparse: $N^2$	fully distributed: $N^4$
Solution of the discrete equations	Use of band struc- ture and right ordering: $(N^2)^{3/2} = N^3$	Gaussian elimination $(N^3)^2 = N^6$
Computation of the solution at all inner grid points	Already known _____	Compute boundary integrals in all grid points of $\Gamma$ , i.e. $N^3$ $N^5$

The above comparison shows rather clearly that the computational expense is in both cases of the same magnitude, i.e. proportional to  $N^3$  or  $N^6$ , respectively. Thus the reduction of one dimension to the boundary integral method is no reduction in computing time. However there are several other properties of boundary integral methods which may be very advantageous:

- (i) The experiments showed very reasonable results already for small numbers of grid points on the boundary  $\Gamma$ .
- (ii) The method is applicable to interior as well as to exterior problems without modifications.

the required potentials are given by boundary potentials and hence can be differentiated analytically away from  $\Gamma$ .

On the other hand the boundary integral method is restricted to problems where the fundamental solution is explicitly available whereas the usual finite element procedures provide a rigorous method.

### 1. Boundary Integral Equations

For the reduction of interior or exterior boundary value problems as well as transmission problems to equivalent boundary integral equations on the boundary manifold  $\Gamma$  one finds many different methods, since this reduction is by no means unique. The two most popular methods are called the "direct method" and the "method of potentials". In all these cases one needs a fundamental solution, respectively, fundamental matrix  $\gamma(z, \zeta)$  of the differential equations explicitly since it will be used in numerical computations. Thus, the practical usefulness of the boundary integral methods hinges essentially on the simple computability of a fundamental solution. This restricts these methods mainly to differential equations with constant coefficients.

For explanation let us consider the exterior plane boundary value problem for the Laplacian in the form

$$(1.1) \quad \begin{aligned} \Delta U &= 0 \quad \text{for } z \in \Omega_e \subset \mathbb{R}^2, \\ U_{\nu} &= f \quad \text{on the boundary } \Gamma, \\ U(z) - B \log |z| &= 0(1) \quad \text{for } |z| \rightarrow \infty. \end{aligned}$$

Here  $\Gamma$  is a simple closed plane curve and  $\mathbb{E}_e$  denotes the exterior domain with boundary  $\Gamma$ . The exterior problem (1.1) describes e.g. a charged conductor in two dimensions [70, p. 174]. The solution of (1.1) can be represented via Green's formula in the form

$$(1.2) \quad U(z) = -\frac{1}{2\pi} \int_{\Gamma} U(\zeta) \frac{\partial}{\partial v_{\zeta}} (\log |z-\zeta|) d\zeta + \frac{1}{2\pi} \int_{\Gamma} \frac{\partial U}{\partial v} (\zeta) \log |z-\zeta| d\zeta + \frac{1}{2}$$

where

$$(1.3) \quad \frac{1}{2\pi} \int \frac{\partial U}{\partial v} d\zeta = B \quad \text{and}$$

$$(1.4) \quad \frac{1}{2} \omega = \lim_{|z| \rightarrow \infty} (U(z) - B \log |z|).$$

Hence, if  $v = \frac{\partial U}{\partial v}|_{\Gamma}$  and  $\omega$  are known then (1.2) gives the solution. The limit  $z \rightarrow \Gamma$  with the jump relation for double layer potentials yields with the "direct method" an integral equation for  $v$  and  $\omega$ ,

$$(1.5) \quad -\frac{1}{\pi} \int_{\Gamma} v(\zeta) \log |z-\zeta| d\zeta - \omega = f$$

$$= \phi - \frac{1}{\pi} \int_{\Gamma} \phi \left( \frac{\partial}{\partial v_{\zeta}} \log |z-\zeta| \right) d\zeta,$$

$$(1.6) \quad \frac{1}{2\pi} \int_{\Gamma} v d\zeta = B.$$

Here  $B$  and  $f$  are given and  $v$  and  $\omega$  are the unknowns. (1.5) is a Fredholm integral equation of the first kind.

For the method of potentials we try to find the solution of (1.1) in the form of a double layer potential,

$$(1.7) \quad U(z) = B \log |z| + \frac{1}{2} \omega + \frac{1}{2\pi} \int_{\Gamma} v(\zeta) \left( \frac{\partial}{\partial \nu} \log |\zeta - z| \right) d\zeta$$

where the double layer density  $v$  and the constant  $\omega$  are to be determined. Since the last potential vanishes for constant  $v$ , we can add the condition

$$(1.8) \quad \frac{1}{2\pi} \int_{\Gamma} v ds = 0.$$

Transition  $z \rightarrow \Gamma$  yields the boundary integral equation

$$(1.9) \quad v(z) - \frac{1}{\pi} \int_{\Gamma} v(\zeta) \left( \frac{\partial}{\partial \nu} \log |\zeta - z| \right) d\zeta - \omega = f(z)$$

$$= 2B \log |z| - 2\phi(z),$$

$$(1.10) \quad \frac{1}{2\pi} \int_{\Gamma} v ds = 0.$$

Here  $f$  is given and  $v$  and  $\omega$  need to be determined.

Similarly, the exterior and interior Neumann problems as well as the interior Dirichlet problem can be formulated in terms of different boundary integral equations.

The above exterior Dirichlet problem is only one very simple example leading to boundary integral equations. References to many other examples can be found in [79], in particular from conformal mapping, electrostatics, flow problems including slow viscous flows, plate and shell problems, elasticity problems in two and three dimensions, punch and crack problems, problems of thermal activity, time harmonic and stationary electromagnetic fields (see also [44] and e.g. the conference proceedings [16, 17, 68]).

In many of these cases the integral equations become much more complicated. However, the types of integral equations are Fredholm integral equations of the second kind as in (1.9) or of the first kind as in (1.5). In addition one also finds singular integral equations on curves as

$$(1.11) \quad a(z)u(z) + \frac{1}{\pi i} b(z) \int_{\Gamma} \frac{u(\zeta)}{\zeta - z} d\zeta + \int_{\Gamma} k(z, \zeta)u(\zeta)ds_{\zeta} = f(z), \quad z \in \Gamma$$

or the corresponding equations on boundary surfaces [54], or one finds operators of the form

$$(1.12) \quad -\frac{\partial}{\partial v_z} \frac{1}{\pi} \int_{\Gamma} u(\zeta) \left( \frac{\partial}{\partial v_{\zeta}} \log |z - \zeta| \right) ds_{\zeta} = f(z), \quad z \in \Gamma,$$

and the corresponding operators in higher dimensions. (e.g. see [37]; in [9] the operator  $K_+$  has (1.12) as principal part.)

Often the above operators also appear in systems of integral equations.

## 2. Strongly Elliptic Integral Equations

Although all the above mentioned types of equations have very different properties in classical theory of integral equations it turns out that if they are considered as so called pseudo-differential operators [76] they have a very strong, common property. Namely the equations of practical interest are "strongly elliptic". In order to formulate this property one needs the Sobolev spaces  $H^s(\Gamma)$  of generalized functions on  $\Gamma$ , their interpolation spaces and their dual spaces. For the definitions we

refer to [3] (in particular p. 214). Then each of the above mentioned operators  $A$  defines a continuous linear mapping  $A : H^s \rightarrow H^{s-2\alpha}$  for a whole scale of real  $s$  (depending on the smoothness of  $\Gamma$ ).  $2\alpha$  is called the order of the pseudodifferential operator  $A$  [76]. (G. Richter calls  $-2\alpha$  in [64] "smoothing index".) For our examples we have  $2\alpha=0$  in (1.9) and (1.11),  $2\alpha = -1$  in (1.5), ~~2α = -1~~ in (1.12). The boundary integral equation we write in short

$$(2.1) \quad Au = f \quad \text{on} \quad \Gamma.$$

The announced common property is the coerciveness in form of the Garding inequality:

$$(2.2) \quad \operatorname{Re}(Av, v) = \operatorname{Re} \int_{\Gamma} v \bar{Av} \, ds \geq \gamma \|v\|_{H^\alpha}^2 - |k[v, v]| \quad \text{for all } v \in H^\alpha(\Gamma)$$

where  $\gamma > 0$  is a constant independent of  $v$  and where  $k[u, v]$  denotes a compact bilinear form on  $H^\alpha \times H^\alpha$ . In some cases  $k$  equals zero, then inequality (2.2) corresponds to strong energy estimates as in [56].

In order to characterize those equations or systems of equations that provide coerciveness let us use the above mentioned context of pseudodifferential operators and let us consider a more general case of systems of equations in the form (2.1). Then to  $A$  there belongs a  $p \times p$  matrix-valued principal symbol  $a_0(x, \xi) = ((a_{q,r}(x, \xi)))_{q,r=1,\dots,p}$  corresponding to the  $p$  equations of (2.1) for the  $p$  components  $v_q$ ,  $q = 1, \dots, p$ . As usual, the  $a_{q,r}(x, \xi)$  are assumed to be homogeneous in  $\xi \in \mathbb{R}^n$  for  $|\xi| \geq 1$  with degrees  $\alpha_{qr} \in \mathbb{R}$ .

Now we define strong ellipticity (analogously to the Agmon-Douglis-Nirenberg ellipticity for differential equations) assuming that there is an index vector  $\alpha = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^p$  such that

$$(2.3) \quad \alpha_{qr} = \alpha_q + \alpha_r, \quad q, r = 1, \dots, p.$$

$A$  is then a continuous linear pseudodifferential operator of order  $2\alpha$ , i.e. defining a continuous map

$$(2.4) \quad A: H^{s+\alpha}(\Gamma) := \bigcap_{q=1}^p W_2^{s+\alpha_q}(\Gamma) \rightarrow H^{s-\alpha}(\Gamma) := \bigcap_{q=1}^p W_2^{s-\alpha_q}(\Gamma), \quad s \in \mathbb{R},$$

in the scale of Sobolev spaces in (2.4). (The admissible  $s$  depend also on the smoothness of  $\Gamma$ .)

Now for the following we assume

$$(2.5) \quad A \text{ is } \underline{\text{strongly elliptic}}$$

i.e. there exists a complex valued smooth matrix  $\theta(z)$  and a constant  $\gamma > 0$  such that

$$(2.6) \quad \operatorname{Re} \zeta^T \theta(z) a_0(z, \xi) \bar{\zeta} \geq \gamma |\zeta|^2$$

for all  $z \in \Gamma$ , all  $\xi \in \mathbb{R}^n$  with  $|\xi|=1$  and for all  $\zeta \in \mathbb{C}^p$ . A strongly elliptic system  $A$  satisfies the Gårding inequality [46],

$$(2.7) \quad \operatorname{Re} (\partial A v, v)_{L^2(\Gamma)} \geq \gamma' \|v\|_{H^\alpha(\Gamma)}^2 - |k[v, v]| \quad \text{for all } v \in H^\alpha(\Gamma)$$

where  $\beta' > 0$  and where  $k[v,w]$  is a compact bilinear form on  $H^{\beta} \times H^{\beta}$ . In the following we shall always consider the equations

$$(2.8) \quad f v := v A + \alpha v =: a$$

instead of (2.1) i.e. for  $A$  we have  $\theta = 1$ . Then Gårding's inequality (2.7) for  $A$  implies the (non unique) decomposition

$$(2.9) \quad A = D + K$$

where  $D$  is a positive definite pseudodifferential operator and  $K : H^{S+1} \rightarrow H^{S+1}$  is compact. This is just the kind of operators that provide the convergence of Galerkin's method for any approximating family of finite dimensional subspaces [77].

In the following table we have collected the main properties of the simple examples of Section 1. Note that the symbols and, hence, strong ellipticity and  $\alpha$  remain the same for equations (1.10), (1.11), (1.12) in higher dimensions if  $|z-w|^{-\frac{1}{n}} \log |z-w|$  is replaced by the corresponding fundamental solution  $\frac{1}{|x-y|^{2-n}}$ , due to the homogeneity. Since for  $D$  and the principal symbol we reflect elliptic operators, we have the same properties if the fundamental solution of the Laplacian is further replaced by that of the field. The equation or by Green's functions to larger domains.

Further examples including symbols of systems can be found in [79, 161].

Equ. (1.9):  $2\alpha = 0$ ,  $Du = u$

$$a_0 = 1$$

, always strongly elliptic

Equ. (1.5):  $2\alpha = -1$ ,  $Du = -\frac{1}{\pi} \int_{\Gamma} \log \left| \frac{z-\zeta}{2\text{diam}\Gamma} \right| u(\zeta) d\zeta$

$$a_0 = \frac{1}{|\zeta|}$$

, always strongly elliptic

Equ. (1.11):  $2\alpha = 0$ ,  $Du = \theta(z)(a(z)u(z) + \frac{b(z)}{1-\bar{z}} \int_{\Gamma} \frac{u(\zeta) d\zeta}{\zeta-z})$

$$a_0 = a(z) + b(z) \frac{\zeta}{|\zeta|}$$

, strongly elliptic iff  
 $a(z) + \lambda b(z) \neq 0$   
 for all  $\lambda \in [-1, 1]$

$\theta(z) = \bar{a} - \bar{b} \lambda_0$  with

$$\lambda_0 := \begin{cases} \operatorname{Re} \bar{a} / |\bar{b}|^2 & \text{for } |\operatorname{Re} \bar{a}| < |\bar{b}|^2 \\ 1 & \text{for } \operatorname{Re} \bar{a} > |\bar{b}|^2 \\ -1 & \text{for } \operatorname{Re} \bar{a} = |\bar{b}|^2 \end{cases}$$

Equ. (1.12):  $2\alpha = 1$ ,  $Du = -\frac{\partial}{\partial \bar{z}} \frac{1}{z} \int_{\Gamma} u(\zeta) \left( \frac{1}{\delta \nabla \zeta} - \log |z-\zeta| \right) ds$

$$a_0 = \dots$$

, always strongly elliptic

$$+ \int u ds$$

Table 1: Positive definite parts, principal symbols and strong ellipticity of the simple examples in Section 1

## IV. The problem of boundary element finite elements

We want to assume that the system of linear boundary integral equations has

$$\text{rank } A = \text{rank } A^T = \text{rank } A^T A = n.$$

It is usually elliptic and has no eigensolution, i.e. the solution is unique and stable. Therefore it is uniquely solvable except for the alternative which is valid because of the ellipticity. If (4.1) admits eigensolutions then finitely many additional linear equations determine the solution uniquely. If the system of nonlinear equations is well then our following theorem remains also completely (see [45, 44, 72, 80]).

Let the following family of finite dimensional approximating spaces be defined by parameters  $n, j$ ,  $j=0, 1, \dots, N$  and the basis  $\phi_{n,j}$ . Then the well known Galerkin procedure provides an element  $\psi_n$  in the  $n$ -dimensional approximation space.

$$(\psi_n, \phi_{n,k})_{\Gamma_1} = (\psi_n, \phi_{n,k})_{\Gamma_2}, \quad k=0, \dots, N,$$

and the following system of linear equations,

$$\sum_{k=0}^N \frac{(\psi_n, \phi_{n,k})_{\Gamma_1}}{(\phi_{n,k}, \phi_{n,k})_{\Gamma_2}} \phi_{n,k} = f_k, \quad k=0, \dots, N.$$

For the convergence of this procedure we have well known results (cf. e.g. C. Johnson [51], S. Hildebrandt and L. Wienhold [20] or the well known lemma known as Céa's lemma, [21, p. 104].

For its formulation let us denote by  $P_h$  the  $L_2$  orthogonal projection onto  $\tilde{H}_h$ . Then we require the approximation property

$$(3.4) \quad \lim_{h \rightarrow 0} \|P_h g - g\|_{H^\alpha} = 0 \text{ for any } g \in H^x.$$

This assumption implies with the Banach-Steinhaus theorem the stability

$$(3.5) \quad \|P_h\|_{H^\alpha, H^\alpha} \leq c \quad \text{for all } h$$

where  $c$  is independent of  $h$ , and also by duality

$$(3.6) \quad \|P_h\|_{H^{-\alpha}, H^{-\alpha}} \leq c.$$

These requirements are satisfied for regular finite elements and also for trigonometric polynomials on closed curves as well as for spherical harmonics on closed boundary manifolds  $\Gamma$  in higher dimensions.

Now we are in the position to state Céa's lemma.

THEOREM 3.1: Let Eq. (3.1) with  $A$  be a strongly elliptic equation with unique solution  $u \in H^0$  to any  $f \in H^{-\alpha}$ . Then there exists  $h_0 > 0$  such that Equations (3.3) are uniquely solvable for every  $0 < h \leq h_0$ . Moreover there exists a constant  $c$  independent of  $h$  and  $f$  such that

$$(3.7) \quad \|\hat{v} - u\|_{H^\alpha} \leq c \inf_{x \in \tilde{H}_h} \|u - x\|_{H^\alpha}$$

For convenience, in the following asymptotic error analysis we are always using  $c, c', \dots$  as generic constants which might change their size and meaning at different places.

As was mentioned above, Theorem 3.1 is not restricted to our finite element approximations but applies to a rather wide class of Galerkin methods as e.g. for the projection methods using trigonometric polynomials as in [61,63].

Proof: Although the proof is standard, let us repeat the main arguments. The Galerkin equations (3.3) are equivalent to finding  $\hat{v} \in \tilde{H}_h$  with

$$(3.8) \quad P_h A P_h \hat{v} = P_h D P_h \hat{v} + P_h K P_h \hat{v} = P_h A u .$$

Since  $D$  is positive definite we have from (2.7) the stability estimate

$$(3.9) \quad \operatorname{Re}(D P_h v, P_h v) \geq \gamma \|P_h v\|_{H^\alpha}^2$$

which yields with the continuity of  $b$  and duality of  $H^{-\alpha}$  and  $H^\alpha$  the stability estimate

$$(3.10) \quad (P_h D P_h)^{-1} : H^{-\alpha}, H^\alpha \rightarrow C$$

on  $H_h$  where  $c$  is independent of  $h$ . Thus we can write

$$(3.11) \quad P_h A P_h = P_h D P_h (I + (P_h D P_h)^{-1} P_h K P_h)$$

The sequence of operators

$$(P_h D P_h)^{-1} P_h K P_h$$

is a composition of inverse stable and, hence, elementwise

convergent operators (3.10),  $P_h$  and the compact operator  $K$ .

Therefore the convergence of

$$\lim_{h \rightarrow 0} \left\{ (P_h D P_h)^{-1} P_h K P_h - D^{-1} K \right\} g = 0$$

for  $g \in H^1$ ,  $\|g\|_{H^1} \leq 1$  is uniform due to [5] and we have

$$\lim_{h \rightarrow 0} (I + (P_h D P_h)^{-1} P_h K P_h)^{-1} g = (I + D^{-1} K)^{-1} g$$

for any  $g \in H^1$  since  $\underbrace{A^{-1} D}_{(I + D^{-1} K)^{-1}} = (I + D^{-1} K)^{-1}$  exists. This implies the uniform boundedness, i.e. stability

$$\|(I + (P_h D P_h)^{-1} P_h K P_h)^{-1}\|_{H^1, H^1} \leq c$$

for all  $h > 0$ ,  $h \leq h_0$  with an appropriate  $h_0 > 0$  where  $c$  is independent of  $h$ . Consequently, an  $H_h$  holds stability

$$(3.12) \quad \|P_h A P_h\|^{-1} = \|(I + (P_h D P_h)^{-1} P_h K P_h)^{-1} (P_h D P_h)^{-1}\| \leq c$$

for all  $0 < h \leq h_0$  where  $c$  is independent of  $h$ . Now (3.12) implies with the continuity of  $A$  the stability of the Galerkin projection.

$$(3.13) \quad G_h = (P_{h\sim} A P_{h\sim})^{-1} P_{h\sim} A, \text{ i.e.}$$

$$(3.14) \quad \|G_h\|_{H^\alpha, H^\alpha} \leq c$$

for all  $0 < h \leq h_0$ . Since  $G_h|_{\tilde{H}_h} = I$  and  $\hat{v} = G_h u$ , Céa's Lemma (3.7) is an immediate consequence. This completes the proof.

Now we specify the spaces  $\tilde{H}_h$  to regular  $(m+1, m)$  systems of finite element functions [11].<sup>1)</sup> They have the following approximation property and satisfy an inverse assumption:

Approximation property:

Let the multiindices  $m, t, s$  satisfy componentwise  $-m-1 \leq t \leq s \leq m+1$ ,  $-m \leq s, t \leq m$ . Then to any  $u \in H^s(\Gamma)$  and any  $h > 0$  there exists a  $\mu \in \tilde{H}_h$  such that

$$(3.15) \quad \|u_q - \mu_q\|_{H^t(\Gamma)} \leq c h^{s_q-t_q} \|u_q\|_{H^{s_q}} \quad (\text{see [15].})$$

The constant  $c$  is independent of  $\mu_q$ ,  $h$  and  $u_q$ .

The finite element functions  $\mu = (\mu_1, \dots, \mu_p) \in \tilde{H}_h$  provide for  $-m \leq t \leq s \leq m$  the inverse assumption

$$(3.16) \quad \|\mu_q\|_{H^{s_q}(\Gamma)} \leq c h^{t_q-s_q} \|\mu_q\|_{H^{t_q}(\Gamma)}$$

where the stability constant  $c$  is independent of  $\mu$  and  $h$  [58].

If we insert (3.15) into the right hand side of (3.14) we surely find improved asymptotic orders of convergence if  $h \rightarrow 0$ . Using

<sup>1)</sup> In general one uses  $(\ell, m)$  systems rather than specifying  $\ell = m + 1$ . We avoid these details here.

the inverse assumption (3.16) one can also extend the estimate of the left hand side to  $H^t$  norms with  $\tau < t \leq m$  [72]. These are the results which have also been obtained with variational methods as in [21]. But as was already mentioned in the introduction, for pseudodifferential operators  $\Lambda$  one can even prove superapproximation [42]. Collecting these results we find the following improved convergence theorem.

THEOREM 3.2 [42, 64, 72]:

Let  $\Lambda$  be strongly elliptic and let (3.1) have an unique solution. Let  $H_h$  satisfy (3.15) and (3.16) and define

$$(3.17) \quad c'_q := \min\{c_q, \alpha\} \quad , \quad q = 1, \dots, p.$$

Suppose  $c_q = c_q(t + s)$ ,  $s \leq c_q - c_q = c_2 - \max\{\alpha, t\} \leq m_q - c_q = c_1$  for  $q = 1, \dots, p$ ,  $s \geq 0$ . Then we have the asymptotic error estimate

$$(3.18) \quad \|u - v\|_{H^{t+s}} \leq c h^{s-t} \|u\|_{H^{t+s}(\Gamma)} .$$

In addition, if we consider the discrete equations (3.3) in  $L_2(\Gamma)$  then we find for the stability of these equations

$$(3.19) \quad \|v\|_{L_2(\Gamma)}^2 \leq c \sum_{q=1}^p h^{2c'_q} \|g\|_{L_2(\Gamma)}^2 .$$

Remarks: With a simple analysis of  $P_h \Lambda P_h = \Lambda$  for  $\omega > 0$  the stability (3.19) yields the conditioning number of (3.3) to be

of the order

$$\sum_{q=1}^p h^{-2|\alpha_q|}.$$

The stability estimate (3.19) can also be used for an estimate of errors due to numerical noise and round off effects in the framework of ill posed problems. This can be found in [43].

The asymptotic estimate (3.18) includes the case  $t < \alpha$ , i.e. superapproximation. If  $t = 2\alpha - m - 1$  then one has for sufficiently smooth data the superapproximation

$$(3.20) \quad \|u - \hat{v}\|_{H^{-m-1+2\alpha}} \leq c h^{2m_q + 2 - 2\alpha_q} \|u\|_{H^{m+1}}.$$

That implies for the desired solution  $\hat{u}$  of the boundary value problems (1.1) inner superconvergence

$$(3.21) \quad \|v - \hat{u}\|_{X(\tilde{\Omega})} \leq c \|u - \hat{v}\|_{H^{-m-1+2\alpha}} \leq c' h^{2m_q + 2 - 2\alpha_q} \|u\|_{H^{m+1}(\Gamma)}$$

where  $\tilde{\Omega}$  is any compact subdomain in the interior, respectively, exterior of  $\Gamma$  and  $X(\tilde{\Omega})$  denotes any norm. Here  $c, c'$  depend on and  $X(\tilde{\Omega})$ .

#### Proof of Theorem 3.2:

Since (3.18) follows for  $\alpha \leq t$  immediately from Céa's Lemma (3.7) with (3.15) and (3.16) let us here indicate the proof only for  $2\alpha - m - 1 \leq t < \alpha$ , i.e. the case of the Aubin-Nitsche Lemma. Moreover let us consider only the case

of one single equation (3.1) instead of a system.

Let us first note that the usual proof of the Aubin-Nitsche Lemma, e.g. [21, p. 137], would yield only  $L_2$ -estimates, i.e.  $t = 0$ . Thus we use a slight modification.

Let us denote by

$$(3.22) \quad e = u - \hat{v}$$

the error term from (3.18). Then (3.3) implies that

$$(3.23) \quad (Ae, \chi)_{L_2} = (e, A^* \chi)_{L_2} = 0 \text{ for all } \chi \in \hat{H}_h.$$

From the existence of  $A^{-1}$  and the strong ellipticity it follows that the adjoint equation

$$(3.24) \quad A^* w = \phi$$

is uniquely solvable for every  $\phi \in H^{-t}(\Gamma)$  with  $w \in H^{2\alpha-t}$ .

Moreover, the continuity of  $A^{*-1}$  implies

$$(3.25) \quad \|w\|_{H^{2\alpha-t}} \leq c \|\phi\|_{H^{-t}}.$$

Since  $H^t(\cdot)$  and  $H^{-t}(\Gamma)$  form a duality with respect to the  $L_2$ -scalar product, we have with (3.24),

$$\begin{aligned} \|e\|_{H^t} &\leq \sup_{\|\phi\|_{H^{-t}} \leq 1} |(e, \phi)| \\ &= \sup |(e, A^* w)| = \sup |(e, A^*(w - \chi)) + (e, A^*\chi)| \end{aligned}$$

$$\begin{aligned} &= \sup |(e, A^*(w - \chi))| \end{aligned}$$

$$\begin{aligned} &\leq \sup_{H^\alpha} \|e\|_{H^\alpha} \|A^*(w - x)\|_{H^{-\alpha}} \\ &\leq c \sup_{H^\alpha} \|e\|_{H^\alpha} \|w - x\|_{H^\alpha} \quad \text{for every } x \in \tilde{H}_h . \end{aligned}$$

Inserting (3.18) for  $t = \alpha$  (that follows from (3.7) with (3.15)) and (3.15) in the above, we find

$$\|e\|_{H^t} \leq \sup c h^{s-\alpha} \|u\|_{H^s} \|w\|_{H^{2\alpha-t}} h^{\alpha-t}$$

if  $-m \leq 2\alpha-t \leq m+1$ , i.e.  $2\alpha-m-1 \leq t \leq m+2\alpha$ . Finally, we use (3.25) to find the desired estimate

$$\|e\|_{H^t} \leq \sup_{\|\phi\|_{H^{-t}} \leq 1} c h^{s-t} \|u\|_{H^s} \|\phi\|_{H^{-t}} = c h^{s-t} \|u\|_{H^s} .$$

#### 4. The Galerkin Collocation Method

For the numerical implementation of Galerkin's procedure (Equations (3.3)), the weights of the influence matrix,

$$(4.1) \quad a_{jk} := (\hat{A}u_j, u_k) , \quad j, k = 0, \dots, N$$

have to be evaluated. Since  $\hat{A}$  is given by an integral operator (in the usual or the generalized) sense, the computation of  $a_{jk}$  requires a double integration over  $\Gamma \times \Gamma$ . If this is done numerically, the kernels of the integral operators must be computed at all combinations of grid points on  $\Gamma$ . In addition, special care must be taken of the singular integrals. In order to reduce the computing time for the evaluation of the stiffness

matrix (4.1) and in order to simplify the computation of the singular integrals let us specify the further investigations to two dimensional problems, i.e.  $\Gamma$  is a plane and---for brevity---closed curve. We further assume that the principal parts of  $A$  are given by convolutional operators. For simplicity let us consider just one equation (3.1). The extension to systems is of simplest technical nature (see [79]). Let  $\Gamma$  be given by a regular parameter representation

$$(4.2) \quad \Gamma: z = z(t), \quad t \in [0,1]$$

with  $z(t)$  an 1-periodic sufficiently smooth vector valued function satisfying

$$(4.3) \quad \left| \frac{dz}{dt} \right| = R(t) \geq R_0 > 0 \text{ for all } t,$$

where  $R$  denotes the Jacobian. Then the operator  $A$  with a convolution operator as principal part has the form

$$(4.4) \quad Au|_t = p.v. \int_{|t-\tau|<\frac{1}{2}} [p_1(t-\tau) + \log|t-\tau| p_2(t-\tau)](u(t)R(t))dt \\ + \int_{|t-\tau|<\frac{1}{2}} L(\tau, t)(u(t)R(t))dt = f(t).$$

Here  $p_1(\zeta)$  and  $p_2(\zeta)$  for  $\zeta \neq 0$  are homogeneous functions of degree  $\beta = -2\alpha-1$ . The principal symbol  $a_0$  and (4.4) are related by the Fourier transformation  $\tilde{F}$ ,

$$(4.5) \quad a_1(\xi) := \tilde{F}(p_1(\cdot) + \log |\cdot| p_2(\cdot))|_{\xi}.$$

For singular integral equations with the Cauchy kernel, the above special form (Equation (4.4)) of A is too restrictive. We leave this detail to [80].

From now on we consider strongly elliptic integral equations of the form of Equation (4.4) and we further assume that the remaining terms collected in  $L(\tau, t)$  define a sufficiently smooth function of  $\tau$  and  $t$ . Otherwise we again split into two terms, where the first contains the singularity and has to be treated similarly to the principal part.

Since in Equation (4.4) only R depends on  $\Gamma$  we consider Equation (4.4) as an integral equation over  $[0,1]$  for the 1-periodic new unknown function

$$(4.6) \quad v(t) := R(t)u(t).$$

Note that the principal part in Equation (4.4) then becomes independent of the special choice of the curve  $\Gamma$ . Therefore we shall adapt numerical integration to the special integrals in Equation (4.4).

The principal part in the standard form (Equation (4.4)) will be handled independently of the special boundary  $\Gamma$  yielding a Toeplitz matrix whose elements are given by a vector. This vector can be computed exactly up to the desired accuracy once for all independent of  $\Gamma$  as well as of  $h$  for any fixed type of element, i.e. shape function. It should be pointed out that

the accuracy of the numerical results depends significantly on how to compute the approximate principal part.

The Galerkin weights due to the smooth remaining parts will be treated numerically by appropriate quadrature formulas depending on the particular finite elements to be used. In them we use only grid points in a regular grid connected with the finite elements such that the kernel functions are to be evaluated as seldom as necessary. This leads to simple modified collocation formulas and the computation of the corresponding stiffness matrix is extremely fast.

In order to utilize the convolution in the principal part we use regular finite elements on a uniform grid of  $[0,1]$  defined with shifts and stretched variables from one shape function  $\mu(\eta)$ . The latter we define as in [8, Chap. 4] by suitable piecewise polynomials of order  $m$  with  $\mu \in C^{m-1}$ .

For  $m = 0, 1, 2$  e.g. we have

$m = 0$	$m = 1$	$m = 2$	for
1	$\eta$	$\frac{1}{2}\eta^2$	$0 \leq \eta < 1$
0	$2 - \eta$	$-\eta^2 + 3\eta - 3/2$	
0	0	$\frac{1}{2}\eta^2 - 3\eta + 9/2$	
0	0	0	elsewhere

With  $\mu$  we define a basis of  $\tilde{H}_h$  by

$$(4.8) \quad u_j(t) := \mu\left(\frac{t}{h} - j\right)\frac{1}{R(t)} \text{ for } h_j \leq t \leq 1+h_j, \quad j=0, \dots, N,$$

$$h=1/(N+1)$$

and their 1-periodic extensions

$$u_j(t + \ell) := u_j(t) \text{ for integer } \ell.$$

For  $u$  in Equation (4.4) we use the approximation

$$(4.9) \quad u_h(t) := \sum_{j=0}^N \gamma_j \mu_j(t) .$$

Remarks:

Our boundary elements have been defined by the transplantation of a regular  $(m+1, m)$  system in the parameter domain onto  $\Gamma$  with the local parameter representation of  $\Gamma$ . For calculations, the integrals will be evaluated by using the local coordinates. In those the finite elements appear as simple functions over the parameter domain. This construction of finite elements on  $\Gamma$  requires that the parameter representation is fully available. For the two-dimensional case this is a sensible requirement. In the space, however, the boundary surface has also to be approximated [55].

For the computations we insert (4.9), (4.8) into Equations (4.1) and we find for the terms due to the first expression in Equation (4.4),

$$\begin{aligned} d_{jk} &= \int_0^1 p.v. \int_{|t-\tau| \leq \frac{1}{2}} [p_1(t-\tau) + \log|t-\tau| p_2(t-\tau)] \mu_j(t) R(t) dt \mu_k(\tau) R(\tau) d\tau \\ &= h^{2+\beta} \left\{ \int_{\tau'=0}^{m+1} p.v. \int_{t'=0}^{m+1} [p_1(t'-\tau' + (j-k)) + p_2 \cdot \log|t'-\tau' + (j-k)|] \right. \\ &\quad \cdot \mu(t') \mu(\tau') dt' d\tau' \\ &\quad \left. + \log h \int_{\tau'=0}^{m+1} p.v. \int_{t'=0}^{m+1} p_2(t'-\tau' + (j-k)) \mu(t') \mu(\tau') dt' d\tau' \right\}, \end{aligned}$$

$$(4.10) \quad d_{jk} = h^{2+\rho} \{ w_{1\rho} + w_{2\rho} \log h \} \text{ with } \rho = j - k \in \mathbb{Z}.$$

Here the two vectors of weights

$$(4.11) \quad w_{1\rho} = \int_{\tau'=0}^{m+1} p.v. \int_{t'=0}^{m+1} [p_1(t'-\tau'+\rho) + p_2 \log |t'-\tau'+\rho|] \times u(t') v(\tau') dt' d\tau'$$

$$(4.12) \quad w_{2\rho} = \int_{\tau'=0}^{m+1} p.v. \int_{t'=0}^{m+1} p_2(t'-\tau'+\rho) u(t') v(\tau') dt' d\tau', \quad \rho \in \mathbb{Z}$$

can be computed once for all independent of  $\tau$  and  $h$ . For more details see [37] and [79]. For all the remaining smooth terms in the Galerkin equations to Equation (4.4) we use numerical integration.

Since in the corresponding integrals

$$(4.13) \quad \int_{\text{supp } u_j} f(t) u_j(t) R(t) dt = h \int_{\sigma=0}^{m+1} f(h(j+\sigma)) u(\sigma) d\sigma$$

the finite element functions appear as factors, the numerical integrations are chosen accordingly to the respective reference function  $u$  such that polynomials  $f$  up to the order  $2M+1$  are integrated exactly. This leads to formulas like

$$(4.14) \quad \int_{\text{supp } u_j} f(t) u_j(t) R(t) dt = h \sum_{\ell=-M}^M b_\ell f(z_{j\ell}) + \tilde{R}$$

where

$$(4.15) \quad z_k := z(h(k + \frac{m+1}{2})) \text{ and } z_{j,\ell} := z(h(j + \frac{m+1}{2} + \gamma_\ell)),$$

$$\ell = -M, \dots, M.$$

are the gridpoints subject to the boundary elements and, correspondingly subject to the integration formula.  $\tilde{R}$  denotes the error term which is of order  $h^{2M+2}$ . The simplest choice  $\gamma_\ell = \ell$  yields  $z_{j\ell} = z_{j+\ell}$  and weights  $b_\ell = b_{-\ell}$  as follows:

	$m = 0$	$m = 1$	$m = 2$
	$b_0 \quad b_1$	$b_0 \quad b_1$	$b_0 \quad b_1$
(4.16) $M = 0 :$	1 0	1 0	1 0
$M = 1 :$	$\frac{11}{12} \quad \frac{1}{24}$	$\frac{5}{6} \quad \frac{1}{12}$	$\frac{3}{4} \quad \frac{1}{8}$

For  $\gamma_\ell = \frac{1}{2}\ell$  and  $M = 2$  one has

	$m = 1$			$m = 2$		
	$b_0$	$b_1$	$b_2$	$b_0$	$b_1$	$b_2$
(4.17)	$\frac{13}{30}$	$\frac{4}{15}$	$\frac{1}{60}$	$\frac{2}{5}$	$\frac{7}{30}$	$\frac{1}{15}$

Instead of (4.17) one often uses Gaussian integration formulas, then  $\gamma_\ell$  correspond to the Gaussian nodal points and (4.14) is modified to

$$(4.18) \quad \int_{\text{supp } \mu_j} f(t) \mu_j(t) R(t) dt = \sum_{j=0}^m \sum_{\ell=-M}^M B_\ell f \mu_j R(\bar{z}_{j+v, \ell}) + \tilde{R}$$

where

$$(4.19) \quad \bar{z}_{j+v, \ell} = z(h(j + \frac{1}{2} + \gamma_\ell)) .$$

and  $B_{-M}, \dots, B_M$  are the Gaussian weights.

Using Formula (4.18) for the smooth terms of the weights in Equation (4.4) we obtain

$$(4.20) \quad \int_{\tau=0}^1 \int_{|\tau-t| \leq \frac{1}{2}} L(\tau, t) \mu_j(t) dt \mu_k(\tau) d\tau = h^2 \sum_{\ell, i=-M}^M b_i b_\ell L(z_{ki}, z_{j\ell}) + \tilde{R}$$

with the error term

$$(4.21) \quad |\tilde{R}| \leq h^{s+2} c \left\{ \max \left| \frac{\partial^s L}{\partial t^s} \right| + \max \left| \frac{\partial^s L}{\partial \tau^s} \right| \right\}, \quad 0 \leq s \leq 2M + 2 .$$

Now we are ready to formulate the Galerkin-collocation equations by using Equations (4.8), (4.10) and (4.20). They read as

$$(4.22) \quad \sum_{j=0}^N a_{hjk} \gamma_j := \sum_{j=0}^N \left\{ h^{2+\beta} (w_{1,\rho}(j,k) + \log h w_{2,\rho}(j,k)) \right. \\ \left. + h^2 \sum_{\ell, i=-M}^M b_i b_\ell L(z_{ki}, z_{j\ell}) \right\} \gamma_j \\ = h \sum_{i=-M}^M b_i f(z_{ki}) =: F_k \quad k=0, \dots, N .$$

For saving computing time, the values of  $L$  and  $f$  at the grid points should be evaluated only once at the beginning and then be stored for further use as to build up the stiffness matrix in Equations (4.22).

This suggests a choice  $\gamma_\ell = \ell$  or  $\frac{1}{2}\ell$  or  $\frac{1}{3}\ell$ , etc., in the numerical integration formulas.

For the asymptotic error due to the Galerkin-collocation we shall use the already established error estimates (Formula (3.18)) for Galerkin's method. To this end we abbreviate the Equations (4.22) by

$$(4.23) \quad \sum_{j=0}^N a_{hjk} \gamma_j = F_k, \quad k = 0, \dots, N$$

as mappings in  $\tilde{H}_h$ . If

$$(4.24) \quad w_h = \sum_{j=0}^N \alpha_j u_j$$

then the mapping  $\tilde{A}_h$  associated with Equation (4.23),

$$(4.25) \quad \sum_{\ell=0}^N \beta_\ell u_\ell = \tilde{A}_h w_h$$

will be defined by the linear equations for the coefficients  $\beta_\ell$ ,

$$(4.26) \quad \sum_{\ell=0}^N \beta_\ell (u_j, u_k) = \sum_{j=0}^N a_{hjk} \alpha_j, \quad k=0, \dots, N.$$

Since the Gram matrix  $(\mu_\ell, \mu_k)$  is regular,  $\tilde{A}_h$  in Equation (4.25) is well defined. Correspondingly we define  $\tilde{F} \in \tilde{H}_h$  by

$$(4.27) \quad (\tilde{F}, \mu_k) = F_k \quad \text{for } k = 0, \dots, N.$$

Then the Galerkin Equations (3.3) and the Galerkin collocation Equations (4.22) take the form

$$(4.28) \quad P_h \tilde{A} P_h \hat{v} = P_h f \quad \text{and} \quad \tilde{A}_h v_h = \tilde{F}, \quad \hat{v}, v_h \in \tilde{H}_h \subset L_2 \cap H^\alpha,$$

respectively. One easily obtains the estimate

$$(4.29) \quad \|\hat{v} - v_h\|_{L_2} \leq \|\tilde{A}_h^{-1}\|_{L_2 L_2} \left\{ \|(\tilde{A}_h - P_h \tilde{A} P_h) \hat{v}\|_{L_2} + \|P_h f - \tilde{F}\|_{L_2} \right\}.$$

This estimate shows clearly that we need estimates for stability, i.e.  $\|\tilde{A}_h^{-1}\|_{L_2 L_2}$ , consistency, i.e.  $\|(\tilde{A}_h - P_h \tilde{A} P_h) \hat{v}\|_{L_2}$  and the truncation error  $\|P_h f - \tilde{F}\|_{L_2}$ . Let us begin with the consistency. With Formula (4.21) one can prove the following:

THEOREM 4.1: Let the weights  $w_{1\rho}, w_{2\rho}$  be accurate to an order  $h^\alpha$  and let  $\left(\frac{\partial}{\partial \tau}\right)^{2M+2} L$  and  $\left(\frac{\partial}{\partial t}\right)^{2M+2} L$  be continuous. Then we have the consistency

$$(4.30) \quad |(\tilde{A}_h \mu, v) - (\tilde{A} \mu, v)| \leq \lambda(h) \|v\|_{L_2} \|v\|_{L_2} \quad \text{for all } v \in \tilde{H}_h$$

where

$$(4.31) \quad \lambda(h) \leq c_1 |\log h| h^{\alpha-2\alpha-1} + c_2 h^{2M+2}.$$

From the estimates (4.30) and (3.19) one easily obtains stability.

THEOREM 4.2: Let the assumptions of Theorem 4.1 be fulfilled and in addition let  $a > 1+2(\alpha-\alpha')$ ,  $M > -\alpha'-1$ . Then we have stability, i.e. there exists  $h_0 > 0$  such that

$$(4.32) \quad \|\tilde{A}_h^{-1}\|_{L_2 L_2} \leq ch^{2\alpha'}$$

where  $c$  is independent of  $h$  for all  $0 < h \leq h_0$ .

Finally, the estimation of the error term  $\tilde{R}$  in Equation (4.14) in connection with Equations (4.27) yields for the truncation error:

Theorem 4.3: For  $F_k = h \sum_{\ell=-M}^M b_\ell f(z_{k\ell})$  in Equations (4.27) there holds

$$(4.33) \quad \|P_h f - \tilde{F}\|_{L_2} \leq ch^\sigma \|f\|_{H^\sigma} \text{ with } 1 \leq \sigma \leq 2M+2.$$

Collecting the foregoing estimates and using Formulae (4.29) and (3.18) we find the following estimates for our Galerkin collocation.

Theorem 4.4: For  $a > m+2+2(\alpha'-\alpha)$  and  $M \geq \frac{m-1}{2} - \alpha'$  we find an error estimate

$$(4.34) \quad \|u - v_h\|_{L_2} \leq ch^s \left\{ \|u\|_{H^s} + \|f\|_{H^{s-2\alpha'}} \right\}$$

with  $1 + 2\alpha' \leq s \leq m + 1$  and  $0 \leq s$ .

For  $a > 2m + 3 - 2\alpha'$  and  $M \geq m - \alpha - \alpha'$  we have even the super approximation

$$(4.35) \quad \|u - v_h\|_{H^t} \leq ch^{s-t} \left\{ \|u\|_{H^s} + \|u\|_{L_2} + \|f\|_{H^{s-t-2\alpha}} \right\}$$

provided  $2\alpha-m-1 \leq t \leq s \leq m+1$ ,  $s-t \geq 1-2\alpha'$ .

### 5. Some Numerical Examples

As we can see from the foregoing error estimates, it seems that the Galerkin collocation (Equations (4.22)) combines the theoretical advantages of Galerkin's method with the practical advantages of the collocation methods. For illustration we present some numerical examples treated in [37, 38, 39, 79] with  $\alpha = -\frac{1}{2} = \alpha'$ . There, the choice  $m=2$ ,  $M=1$  and  $\gamma_\alpha=1$  provided excellent numerical results in combination with short computing times.

The boundary integral equations treated so far numerically are all of the form

$$(5.1) \quad - \int_{\Gamma} \log(z-\zeta) \vec{u}(\zeta) d\zeta + \int_{\Gamma} L(z, \zeta) \vec{u}(\zeta) d\zeta = \vec{f}(z) + \vec{\omega},$$

$$\int_{\Gamma} \vec{u}(\zeta) d\zeta = \vec{B}, \quad z, \zeta \in \Gamma.$$

Here  $\vec{f}$  is a given  $n$ -component vector function on  $\Gamma$ ,  $n = 1, 2$ ,  $\vec{B} \in \mathbb{R}^n$  is a given constant vector and  $\vec{u}$  and  $\vec{\omega}$  are the unknown  $n$ -component vector function, respectively, constant vector.  $L$  is a given smooth  $n \times n$  matrix function on  $\Gamma \times \Gamma$ .

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

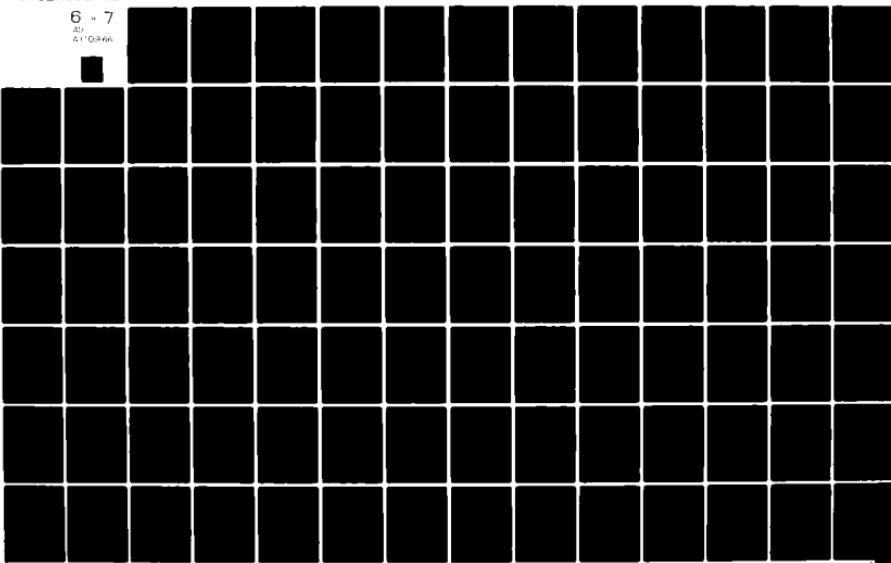
F/6 12/1

UNCLASSIFIED

AFOSR-TR-82-0047

NL

6 x 7  
40  
41-0466



Example 5.1:Symm's method in conformal mapping [74,75,39,79]5.1: Interior Conformal Mapping

Let  $w$  denote the conformal mapping of  $\Omega_i$  onto the unit disc and let  $\theta_i = \arg w|_{\Gamma}$  denote the angle of the boundary mapping. Then Gaier [30] showed that Symm's integral equation [74] for the interior mapping function provides  $\theta_i^*$  as the solution.

Then the slightly modified equations

$$(5.2) \quad - \int_{\Gamma} \log|z-\zeta| u(\zeta) dt_{\zeta} + \omega = -\log|z|, \quad z \in \Gamma,$$

$$\int_{\Gamma} u dt = 1$$

have a unique solution  $u = u(t)$ ,  $\omega$  [40] and with Theorem 12 in [30] it can easily be shown that the unique solution is given by

$$(5.3) \quad u = \frac{1}{2\pi} \frac{d\theta_i}{dt}, \quad \omega = 0,$$

no matter whether the capacity of  $\Gamma$  is 1 or not.

Since the  $\mu_j(t) = \mu(\frac{t}{h} - j) \frac{1}{2\pi}$  are piecewise polynomials, the integrals can be evaluated exactly either with explicit integration or with appropriate most simple numerical formulas. For details see [39].

In the tables we compare the results of our computations with the exact values for three examples of inner mappings in [29].

.. 7 ..

### Interior mapping of ellipses

(See [29, p. 264, Example 3 and p. 161, Table 14a])

$\Gamma: z(t) = (\cos 2\pi t, \delta \sin 2\pi t)$ ,  $0 < \delta \leq 1$ . Computations for  $\delta = 0.2, 0.5, 0.8\bar{3}$  with  $m=2$  and 60 grid points on  $\Gamma$  in double precision (14 decimal digits) showed the following absolute errors:

$\delta$	0.2	0.5	$0.8\bar{3}$
abs. errors	$4 \times 10^{-3}$	$7 \times 10^{-5}$	$3 \times 10^{-6}$

### Interior mapping of reflected ellipses

(See [29, p. 264, Example 2 and pp. 102, 103])

Boundary  $\Gamma$ :  $z(t) = (\cos 2\pi t, \delta \sin 2\pi t)/\{\cos^2 2\pi t + \delta^2 \sin^2 2\pi t\}$

Computing time for each case: 1.3 sec. CPU.

Number of grids point:  $N+1 = 36$

Computations for  $\delta = 0.25, 0.6$  and  $0.65$  with  $m=2$ , and in single precision (7 decimal digits) showed the following absolute errors:

$\delta$	0.25	0.6	0.65
abs. errors	$3 \times 10^{-3}$	$7 \times 10^{-5}$	$10^{-4}$

### Interior mapping of an excentric circle

(See [29, p. 264, Example 1])

Boundary  $\Gamma$ :  $z(t) = e^{i2\pi t} \left( \cos 2\pi t + \sqrt{b^2 - \sin^2 2\pi t} \right)$

Computing time for each case: 4 sec. CPU.

Number of grid points:  $N+1 = 60$ .

Computations for  $b = 5$  and  $5/3$  with  $m = 2$  and in double precision (14 decimal digits) showed the following maximal absolute errors:

b	5	$5/3$
abs. errors	$10^{-7}$	$10^{-6}$

### 5.1.1 Exterior Conformal Mapping:

Here we compute the conformal mapping  $w$  of the exterior domain  $\Omega_e$  onto the exterior of the unit disc and again we are interested in the boundary map given by  $\theta_e = \arg w|_{\Gamma}$ . According to Symm [74] and Gaier [30] we now solve the again modified equations

$$(5.4) \quad - \int_{\Gamma} \log|z - \zeta| u d\zeta - \omega = 0, \quad z \in \Gamma.$$

$$\int u dt = 1.$$

Due to [40] they have a unique solution  $u(t)$ ,  $\omega$ .

With Theorem 11 in [30] it immediately follows that  $u(t)$  and  $\omega$  are given by

$$(5.5) \quad u(t) = \frac{1}{2\pi} \frac{d\theta_e}{dt} \text{ and } \omega = -\log(\text{capacity of } \Gamma) = (\text{Robin's constant}).$$

Hence, the solution of (5.4) provides at the same time the boundary mapping of the exterior mapping and Robin's constant.

We have computed one exterior mapping of an ellipse. (See [29, p. 264, Example 3]). There the boundary curve  $\Gamma$  is chosen by

$$z(t) = 2(\sqrt{3}) \cos 2\pi t + (i/\sqrt{3}) \sin 2\pi t.$$

We chose  $m=2$ ,  $M=1$ ,  $N+1 = 40$  grid points and double precision (14 decimal digits). The boundary mapping is in this case explicitly known as  $\theta_e(t) = 2\pi t$ . The numerical results are accurate up to 10 digits. The computed capacity is

$$\text{capacity } (\Gamma) = 0.8660253881.$$

Example 5.2:

Exterior boundary value problem for the Bilaplacian, the Stokes problem

Here the underlying boundary value problem is the exterior Stokes problem

$$\Delta^2 U = 0 \text{ in } \Omega_e ,$$

$$\nabla U = 0 \text{ on } \Gamma$$

$$\text{and } \nabla U \rightarrow (0, -1) \text{ for } |z| \rightarrow \infty .$$

According to [40] we have the solution

$$U(z) = -\frac{1}{2} \int_{\Gamma} (\nabla_{\zeta} |z-\zeta|^2 \log |z-\zeta|) \cdot (u_1(\zeta), u_2(\zeta)) d\zeta - x\omega_1 - y\omega_2$$

where  $u_1, u_2$  solve the system (5.1) with

$$(5.6) \quad L_{\alpha, \beta} = -\frac{1}{2} \delta_{\alpha \beta} - \frac{(x_{\alpha} - \xi_{\alpha})(x_{\beta} - \xi_{\beta})}{|z - \zeta|^2}$$

For  $\Gamma$  we again choose the ellipses

$$\Gamma: z(t) = (\cos 2\pi t, \delta \sin 2\pi t) .$$

Computations for  $\delta = 0.6, 0.9$  with  $m = 0, 1, 2$  and 20 and 40 grid points on  $\Gamma$  in double precision (14 decimal digits) showed the following absolute errors for  $u_2$ :

$\delta$	0.6		0.9	
$h$	1/20	1/40	1/20	1/40
$m=0$	$10^{-3}$	$2 \times 10^{-4}$	$10^{-3}$	$2 \times 10^{-5}$
$=1$	$2 \times 10^{-6}$	$10^{-7}$	$10^{-5}$	$10^{-7}$
$=2$	$2 \times 10^{-6}$	$10^{-7}$	$10^{-5}$	$10^{-7}$

More details and a further example can be found in [39].

## II Integral Equation Methods for Mixed Boundary Value Problems

This lecture gives a survey on joint work by M. Costabel, G. C. Hsiao, U. Lamp, T. Schleicher, E. Stephan and W. L. Wendland [24, 25, 49, 50, 71, 82, 83].

### Introduction

The application of the boundary element method in the form of Galerkin collocation to mixed boundary value problems requires some modifications. This is due to the singularities of the solution's gradient at the collision points in two dimensional problems and, respectively, at the collision curve in three dimensional problems where the two different boundary conditions are adjoining.

Since Fichera's fundamental work on the Zaremba problem [28], it is well known that these singularities are unavoidable unless the data satisfy specific side conditions. These singularities generate corresponding singularities of the boundary charges in the boundary integral method. They pollute numerical computations unless they are handled separately. Here we shall show how the boundary integral method can be improved by augmenting the appropriate singularity functions to the finite element scheme. This is based on a local analysis of the solution to the mixed boundary value problem due to Grisvard [33] and cf the integral equations [24, 25, 83].

Then we apply Galerkin's method to the modified integral equations. A similar method but with collocation has been used by J. Blue [14]. Since our system of integral equations is strongly elliptic in the sense of (I.2.5 ff.) we find convergence in the corresponding energy norm. This estimate corresponds to [23] and [27]. In order to improve the convergence of the approximation we use local analysis for better regularity in connection with modified coerciveness on one hand and a priori estimates for the corresponding pseudo differential operators on the other hand. We find improved asymptotic convergence and also super approximation which cannot be obtained by variational methods and coerciveness alone. Moreover we approximate the stress intensity factors besides the desired charges and give corresponding error estimates.

In this lecture we shall restrict our presentation mainly to a review on the case of the mixed boundary value problem in a smooth domain following [83] and the corresponding Galerkin collocation which has been worked out in [49,50]. The generalization of the whole method to polygonal domains is here only sketched. It involves much deeper analysis and will be presented in [24,25]. Eventually we shall indicate a formulation of a system of boundary integral equations that governs a three-dimensional mixed boundary value problem [82].

Mixed boundary value problems in two and three dimensions describe many problems of classical mathematical physics as crack and punch problems, contact problems in thermoelasticity, heat conduction in space science, electrostatics and flow and infiltration problems---to name a few. Some of these examples can be found in [69,83].

References to the first lecture are denoted by  
(II.1) etc.

### §1 The Plane Mixed Problem

Let us consider the plane mixed problem with the Laplacian,

$$(1.1) \quad \Delta U = 0 \text{ in } \Omega \subset \mathbb{R}^2 \text{ (or in } \mathbb{R}^2 \setminus \bar{\Omega})$$

$$U = g_1 \text{ on } \Gamma_1 ,$$

$$\frac{\partial U}{\partial \nu} = g_2 \text{ on } \Gamma_2 ,$$

(and an appropriate condition at infinity for exterior problems).  $\Omega$  is a simple connected bounded domain in  $\mathbb{R}^2$  with a smooth boundary curve  $\Gamma = \Gamma_1 \cup \Gamma_2 \cup z_1 \cup z_2$  where  $\Gamma_1$  and  $\Gamma_2$  are two (for simplicity) disjoint parts of  $\Gamma$  with endpoints  $z_1$  and  $z_2$ . For brevity we restrict us to the case of interior mixed problems; the appropriate modifications for exterior problems are easily formulated. We omit the details.

As in (II.1) (II.2) we formulate the boundary integral equations via the "direct method," i.e. via the Green formula with the fundamental solution representing the variational solution  $U$  within the domain  $\Omega$  by

$$\begin{aligned} U(z) &= \frac{1}{2\pi} \int_{\Gamma} U(\zeta) \frac{\partial}{\partial v_\zeta} (\log |z-\zeta|) ds_\zeta \\ (1.2) \quad & - \frac{1}{2\pi} \int_{\Gamma} \frac{\partial U}{\partial v} (\zeta) \log |z-\zeta| ds_\zeta . \end{aligned}$$

Here  $s_\zeta$  denotes the arc length at  $\zeta \in \Gamma$  and  $\frac{\partial}{\partial v_\zeta}$  denotes the normal derivative at  $\zeta \in \Gamma$  in direction of the exterior normal. Replacing  $U$  on  $\Gamma_1$  by  $g_1$  and  $\frac{\partial U}{\partial v}$  on  $\Gamma_2$  by  $g_2$  and passing  $z$  to the boundary  $\Gamma$ , one obtains with the well known jump relations for the double layer potential the following equations on the corresponding parts of the boundary:

$$\begin{aligned} \text{on } \Gamma_2: A_1(U|_{\Gamma_2}, \frac{\partial U}{\partial v}|_{\Gamma_1}) &:= U(z) - \frac{1}{\pi} \int_{\Gamma_2} U(\zeta) \left( \frac{\partial}{\partial v_\zeta} \log |z-\zeta| \right) ds_\zeta \\ (1.3) \quad & + \frac{1}{\pi} \int_{\Gamma_1} \left( \frac{\partial U}{\partial v}(\zeta) \right) \log |z-\zeta| ds_\zeta \\ & = \frac{1}{\pi} \int_{\Gamma_1} g_1(\zeta) \left( \frac{\partial}{\partial v_\zeta} \log |z-\zeta| \right) ds_\zeta \\ & - \frac{1}{\pi} \int_{\Gamma_2} g_2(\zeta) \log |z-\zeta| ds_\zeta , \end{aligned}$$

$$\begin{aligned}
 \text{on } \Gamma_1: A_2(U|_{\Gamma_2}, \frac{\partial U}{\partial v}|_{\Gamma_1}) &:= -\frac{1}{\pi} \int_{\Gamma_1} \left( \frac{\partial U}{\partial v}(\zeta) \right) \log |z-\zeta| \, ds_{\zeta} \\
 &\quad + \frac{1}{\pi} \int_{\Gamma_2} U(\zeta) \left( \frac{\partial}{\partial v} \log |z-\zeta| \right) \, ds_{\zeta} \\
 (1.4) \qquad &= g_1(z) - \frac{1}{\pi} \int_{\Gamma_1} g_1(\zeta) \frac{\partial}{\partial v} (\log |z-\zeta|) \, ds_{\zeta} \\
 &\quad + \frac{1}{\pi} \int_{\Gamma_2} g_2(\zeta) \log |z-\zeta| \, ds_{\zeta}.
 \end{aligned}$$

These two equations now serve as integral equations for the unknown boundary data  $U|_{\Gamma_2}$  and  $\frac{\partial U}{\partial v}|_{\Gamma_1}$ . As soon as these are known (1.2) gives the desired solution in the whole of  $\Omega$ . The analysis of the integral equations (1.3) (1.4) will be presented in the following.

The validity of the above steps must be justified and depends on the behaviour and regularity of  $U$  and  $\frac{\partial U}{\partial v}$  at the boundary. To this end and for the further analysis we need also Sobolev spaces on the boundary parts  $\Gamma_j$  defined as

$$(1.5) \quad H^r(\Gamma_j) := \{f = F|_{\Gamma_j} \text{ with } F \in H^r(\Gamma) \text{ and}$$

$$\|f\|_{H^r(\Gamma_j)} := \inf_F \|F\|_{H^r(\Gamma)}$$

and

$$\tilde{H}^r(\Gamma_j) := \{f \in H^r(\Gamma) \text{ with } \text{supp } f \subseteq \bar{\Gamma}_j\}$$

(1.6)

$$\text{and } \|f\|_{\tilde{H}^r(\Gamma_j)} := \|f\|_{H^r(\Gamma)} .$$

Now let us assume that the data are given with

$$(1.7) \quad g_1 \in H^{(3/2)+\sigma}(\Gamma_1) \text{ and } g_2 \in H^{(1/2)+\sigma}(\Gamma_2), \quad |\sigma| < \frac{1}{2} .$$

Then for the boundary value problem (1.1) we have the following theorem.

Theorem 1.1 [83]: To every  $g_1 \in H^{(3/2)+\sigma}(\Gamma_1)$ ,  $g_2 \in H^{(1/2)+\sigma}(\Gamma_2)$  with  $|\sigma| < 1/2$  there exists exactly one solution  $u$  of the mixed boundary value problem (1.1) of the form

$$(1.8) \quad u(z) = \sum_{i=1}^2 \alpha_i \rho_i^{1/2} \sin \frac{1}{2} \theta_i + v(z)$$

with a smooth function  $v \in H^{2+\sigma}(\Omega)$ ,  $|\sigma| < \frac{1}{2}$ .

Here  $H^{2+\sigma}(\Omega)$  denotes the Sobolev space over the domain  $\Omega$ ,

$\rho_i = |z - z_i|$  denote the distances to the corresponding collision points  $z_i$  and  $\theta_i$  denote the respective angles between the tangent vectors at  $z_i$  in the direction of  $\Gamma_1$  and the rays  $z - z_i$ .

The special form (1.8) of the solution provides the validity of the Green theorem (1.2) and the jump relations (see references in [83]). According to (1.8)

the desired quantities in (1.3), (1.4) can be written as

$$(1.9) \quad U|_{\Gamma_2} = \sum_{i=1}^2 \alpha_i \rho_i^{(1/2)} \chi_i + \tilde{g}_1 + w_o$$

and

$$(1.10) \quad \frac{\partial U}{\partial v}|_{\Gamma_1} = -\frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{(-1/2)} \chi_i + \tilde{g}_2 + \phi_o$$

where  $\tilde{g}_1 \in H^2(\Gamma_2)$  and  $\tilde{g}_2 \in H^1(\Gamma_1)$  are arbitrarily chosen functions satisfying the transition conditions

$$(1.11) \quad \tilde{g}_1(z_i) = g_1(z_i) \quad \text{for } i = 1, 2 ,$$

and

$$\tilde{g}_2(z_i) = g_2(z_i) \quad \text{if } 0 < \sigma < \frac{1}{2} \quad \text{or}$$

$$(1.12) \quad g_2^x(z) := \begin{cases} g_2(z) & \text{for } z \in \Gamma_2 , \\ \tilde{g}_2(z) & \text{for } z \in \Gamma_1 \end{cases} \in H^{(1/2)+\sigma}(\Gamma)$$

$$\text{if } -\frac{1}{2} < \sigma \leq 0 , \quad i = 1, 2 .$$

Then these functions  $w_o \in H^{(3/2)+\sigma}(\Gamma_2) \cap H^1(\Gamma_2)$  and  $\phi_o \in H^{(1/2)+\sigma}(\Gamma_1)$  represent new unknown smooth densities whereas  $\alpha_i$ ,  $i = 1, 2$  are the unknown stress intensity factors.  $\chi_i$ ,  $i = 1, 2$  are two cut-off functions with  $\chi_i \equiv 1$  in some neighborhood of  $z_i$  which will be specified later on.

Since the system (1.3), (1.4) admits an eigen-solution if  $\Gamma$  has conformal radius 1 we further enforce the compatibility condition

$$\int_{\Gamma} \frac{\partial U}{\partial v} ds = \int_{\Gamma_1} \frac{\partial U}{\partial v} ds + \int_{\Gamma_2} g_2 ds = 0$$

as an additional equation whilst introducing a real constant  $\omega \in \mathbb{R}$  as a new unknown that must vanish for the solution of (1.3), (1.4). Inserting (1.9), (1.10) and incorporating the preceding remarks, we find for the system (1.3), (1.4) the final form

$$\begin{aligned}
 (I - K_{22})w_O + R_{12}(\alpha_i, \phi_O) &:= w_O(z) = \frac{1}{\pi} \int_{\Gamma_2} w_O(\zeta) d\theta_z \\
 &\quad + \frac{1}{\pi} \int_{\Gamma_1} \phi_O(\zeta) \log|z-\zeta| ds_\zeta \\
 &\quad + \sum_{i=1}^2 \alpha_i \left[ \rho_i^{1/2} \chi_i - \frac{1}{\pi} \int_{\Gamma_2} \rho_i^{1/2} \chi_i d\theta \right. \\
 &\quad \left. - \frac{1}{2\pi} \int_{\Gamma_1} \rho_i^{-1/2} \chi_i \log|z-\zeta| ds_\zeta \right] \\
 (1.13) \quad &= \frac{1}{\pi} \int_{\Gamma_1} g_1 d\theta + \frac{1}{\pi} \int_{\Gamma_2} \tilde{g}_1 d\theta - \tilde{g}_1(z) \\
 &\quad - \frac{1}{\pi} \int_{\Gamma} g_2^x \log|z-\zeta| ds_\zeta - \omega \\
 &= F_1(z) - \omega \text{ for } z \in \Gamma_2 ,
 \end{aligned}$$

$$\begin{aligned}
v_{11} \left[ \phi_o - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} \chi_i \right] + K_{21} (w_o + \sum_{i=1}^2 \alpha_i \rho_i^{1/2} \chi_i) &:= \\
- \frac{1}{\pi} \int_{\Gamma_1} \left[ \phi_o(\zeta) - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} \chi_i(\zeta) \right] \log |z-\zeta| \, ds_\zeta \\
+ \frac{1}{\pi} \int_{\Gamma_2} w_o \, d\theta + \sum_{i=1}^2 \alpha_i \frac{1}{\pi} \int_{\Gamma_2} \rho_i^{1/2} \chi_i \, d\theta \\
(1.14) \quad &= g_1(z) - \frac{1}{\pi} \int_{\Gamma_1} g_1 \, d\theta - \frac{1}{\pi} \int_{\Gamma_2} \tilde{g}_1 \, d\theta \\
&+ \frac{1}{\pi} \int_{\Gamma} g_2^x \log |z-\zeta| \, ds_\zeta + \omega \\
&= F_2(z) + \omega \quad \text{for } z \in \Gamma_1
\end{aligned}$$

and

$$(1.15) \quad \int_{\Gamma_1} \left[ \phi_o - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} \chi_i \right] \, ds = - \int_{\Gamma} g_2^x \, ds = B .$$

In (1.13), (1.14) we denote by  $d\theta$  the kernel of the double layer potential,

$$(1.16) \quad d\theta_z(\zeta) := \frac{\partial}{\partial v_\zeta} (\log |z-\zeta|) \, ds_\zeta .$$

Note that  $d\theta$  is the total differential of the angle  $\arg(z-\zeta)$ .

The mapping properties of the system of integral equations (1.13) - (1.15) are essentially based on the mapping properties of the logarithmic integral operator on a part of  $\Gamma$ , namely on  $\Gamma_1$ . We find that this operator is bijective in suitable pairs of function spaces.

Besides the explicit knowledge of the exceptional functions we shall also need properties of the logarithmic integral operators on other parts of  $\Gamma$ . To this end we define

$$(1.17) \quad V\psi(z) := -\frac{1}{\pi} \int_{\Gamma} \psi \log|z-\zeta| \, ds_{\zeta} \quad \text{for } z \in \Gamma$$

and

$$(1.18) \quad V_{jk}\psi(\zeta) := -\frac{1}{\pi} \int_{\Gamma_j} \psi \log|z-\zeta| \, ds_{\zeta} \quad \text{for } z \in \Gamma_k; \quad j, k = 1, 2.$$

For  $V_{11}$  we already have the following coerciveness inequality [41].

Lemma 4.1: There exists a constant  $v > 0$  such that

$$(1.19) \quad (V_{11}\psi, \psi)_{L^2(\Gamma_1)} \geq v \|\psi\|_{H^{-1/2}(\Gamma_1)}^2$$

holds for every  $\psi \in H^{-1/2}(\Gamma_1)$ .

We further need the mapping properties of  $V_{jk}$  applied to the exceptional functions

$$(1.20) \quad u_i = \rho_i^{1/2} \sin \frac{1}{2} \theta_i \quad , \quad i = 1, 2 .$$

Explicit calculations with the harmonic functions (1.20)  
yield the following lemma [83, Lemma A.4]:

Lemma 4.2: Let  $\mu < 1$ . With  $u_1$  in (1.20) and  $\frac{1}{2} \rho_1^{-1/2} x_1$   
let us define

$$\begin{aligned} \psi_1 &:= \begin{cases} \frac{1}{2} \rho_1^{-1/2} x_1 + \frac{\partial u_1}{\partial v} (1 - x_2) & \text{on } \Gamma_1 , \\ 0 & \text{on } \Gamma_2 , \end{cases} \\ \psi_2 &:= \begin{cases} 0 & \text{on } \Gamma_1 , \\ \frac{\partial u_1}{\partial v} (1 - x_2) & \text{on } \Gamma_2 . \end{cases} \end{aligned}$$

Then  $\psi_1, \psi_2 \in H^\mu(\Gamma)$ . Furthermore

$$(1.21) \quad v_{1i} \left( \frac{1}{2} \rho_1^{-1/2} x_1 + \frac{\partial u_1}{\partial v} (1 - x_2) \right) \in H^{1+\mu}(\Gamma_i), \quad i = 1, 2$$

$$(1.22) \quad v_{2i} \left( \frac{\partial u_1}{\partial v} (1 - x_2) \right) \in H^{1+\mu}(\Gamma_i), \quad i = 1, 2 \text{ and}$$

$$(1.23) \quad v_{11} (x_1 \rho_1^{-1/2}) \in H^{1+\mu}(\Gamma_1) .$$

The same properties hold for  $u_2$  and  $\frac{1}{2} \rho_2^{-1/2} x_2$  correspondingly.

With the preceding preliminary results we prove in [83] the  
following theorem.

Theorem 4.2: Let us assume that diameter  $(\Gamma) < 1$ . For  $|\sigma| < 1/2$  let

$$z^{(1/2)+\sigma}(\Gamma_1) := \left\{ \{\alpha_1, \alpha_2, \psi_0\} : \alpha_1, \alpha_2 \in \mathbb{R} \text{ and} \right.$$

$$\left. \psi_0 \in H^{(1/2)+\sigma}(\Gamma_1) \right\} .$$

Then the mapping  $v_{11} : z^{(1/2)+\sigma}(\Gamma_1) \rightarrow H^{(3/2)+\sigma}(\Gamma_1)$  with

$$(1.24) \quad \{\alpha_1, \alpha_2, \psi_0\} \mapsto v_{11}\{\alpha_1, \alpha_2, \psi_0\} := v_{11} \left\{ -\frac{\alpha_1}{2} \rho_1^{-1/2} x_1 \right. \\ \left. - \frac{\alpha_2}{2} \rho_2^{-1/2} x_2 + \psi_0 \right\}$$

is bijective and continuous. Moreover, for  $-2 < s < 0$ ,  $s \neq -1$ , the mapping  $v_{11} : H^s(\Gamma_1) \rightarrow H^{s+1}(\Gamma_1)$  is also continuous and bijective.

This theorem enables us to apply the approach of I§2 to our more general situation. Since  $v_{12}$  is continuous but not compact, the principal symbol of (1.13), (1.14) has the form

$$(1.25) \quad a_0 = \begin{pmatrix} 1, & a_{12} \\ 0, & |\xi|^{-1} \end{pmatrix} \quad \text{with } \alpha = (0, -1/2) .$$

Now it is easily seen that  $a_0$  (1.25) is strongly elliptic (see (I2.6)) since for  $\kappa \geq 1+4 \max |a_{12}(z, \xi)|$

$$\frac{|\xi|}{z \in \Gamma} = 1,$$

one finds the inequality

$$(1.26) \quad \operatorname{Re} \left\{ \zeta^T \begin{pmatrix} 1 & 0 \\ 0 & \kappa \end{pmatrix} a_0 \bar{\zeta} \right\} = \operatorname{Re} \left\{ |\zeta_1|^2 + |\zeta_2|^2 + a_{12}(z, \xi) \bar{\zeta}_1 \zeta_2 \right\}$$

$$\geq \frac{3}{4}(|\zeta_1|^2 + |\zeta_2|^2) \text{ for } |\zeta| = 1 \text{ and } \zeta \in \mathbb{S}^2$$

As in (I2.7) we find coerciveness.

Lemma 1.3: To  $A_1, A_2$  (1.3), (1.4) and the above choice of  $\kappa$  there exists a constant  $\gamma_0 > 0$  such that the Gårding inequality

$$(1.27) \quad \begin{aligned} & (A_1(U, \psi), U)_{L_2(\Gamma_2)} + (A_2(U, \psi), \psi)_{L_2(\Gamma_1)} \\ & \geq \gamma_0 \left\{ \|U\|_{L_2(\Gamma_2)}^2 + \|\psi\|_{H^{-1/2}(\Gamma_1)}^2 \right\} - |k((U, \psi), (U, \psi))| \end{aligned}$$

holds for all  $(U, \psi) \in L_2(\Gamma_2) \times H^{-1/2}(\Gamma_1)$  where  $k$  is a suitable compact bilinear form on  $L_2(\Gamma_2) \times H^{-1/2}(\Gamma_1)$ .

As we have seen in (I§3), the coerciveness (1.27) provides Céa's lemma, Theorem I.3.1 and Theorem I.3.2 for the immediate Galerkin approximation of (1.3), (1.4) with finite elements. According to the smoothness of  $U|_{\Gamma_2}$  (1.9) and  $\frac{\partial U}{\partial v}|_{\Gamma_1}$  (1.10) we find the following lemma corresponding to [23] and [27].

Lemma 1.4: Let  $u_h, \psi_h$  denote the Galerkin solutions with regular finite elements,  $m \geq 0$ , to (1.3), (1.4). Then one finds asymptotic convergence as

$$(1.28) \quad \|U - u_h\|_{L_2(\Gamma_2)} + \left\| \frac{\partial U}{\partial v} - \psi_h \right\|_{H^{t+\epsilon}} \\ \leq c_\epsilon \left\{ h^{1-\epsilon} \|U\|_{H^{1-\epsilon}(\Gamma_2)} + h^{-(t+2\epsilon)} \left\| \frac{\partial U}{\partial v} \right\|_{H^{-\epsilon}(\Gamma_1)} \right\}$$

with any  $\epsilon > 0$  and  $-1 \leq t \leq -1/2$ . The constant  $c_\epsilon$  is independent of  $U, h, u_h$ , and  $\psi_h$  but may depend on  $\epsilon$ .

Without special treatment of the singularities this estimate cannot be improved. Thus we need to use a finer analysis of the integral equations (1.13) - (1.15). Defining the space of new unknowns by

$$(1.29) \quad \tilde{w}^{(1/2)+\sigma} = \left\{ \begin{array}{l} \{\alpha_i, \omega, \phi_0, w_0\} \mid \alpha_i \in \mathbb{R}, \omega \in \mathbb{R}, \phi_0 \in H^{(1/2)+\sigma}(\Gamma_1), \\ w_0 \in H^{(3/2)+\sigma}(\Gamma_1), i=1,2 \end{array} \right\}$$

we have the following theorem [83, Theorem 2.3 and Theorem 2.4].

Theorem 1.3: The mapping defined by the left hand sides of (1.13) - (1.15) is an isomorphism

$$\tilde{w}^{(1/2)+\sigma} \rightarrow H^{(3/2)+\sigma}(\Gamma_2) \times H^{(3/2)+\sigma}(\Gamma_1) \times \mathbb{R} =: \tilde{F}^{(3/2)+\sigma}$$

for any  $|\sigma| < 1/2$ .

The proof in [83] is rather involved using (1.27), Fredholm theory and classical potential theory.

Restricting Theorem 1.3 to subspaces one easily finds the following theorem:

Theorem 1.4: (1.13) - (1.15) defines an isomorphism

$$\underline{w}_o^{(1/2)+\sigma} \rightarrow \underline{F}_o^{(3/2)+\sigma} \quad \text{where}$$

$$(1.30) \quad \underline{w}_o := \left\{ \{\alpha_i, \omega, \phi_o, w_o\} \in \underline{W} \mid w_o(z_i) = 0, i = 1, 2 \right\},$$

$$(1.31) \quad \underline{F}_o := \left\{ \{F_1, F_2, B\} \in \underline{F} \mid F_1(z_i) + F_2(z_i) = 0 \right\}.$$

In order to apply the Aubin-Nitsche lemma to the system of integral equations (1.13) - (1.15) one needs a formulation which takes care of the stress intensity factors also in the case that the singularity functions are contained in the respective Sobolev space. To this end we multiply equation (1.14) by  $v_{11}^{-1}$  assuming (without loss of generality) that diameter  $(\Gamma) < 1$ . Then the equations (1.13) - (1.15) take the form

$$(1.32) \quad (I - K_{22})w_o + R_{12}\{\alpha_i, \phi_o\} = F_1 - \omega,$$

$$(1.33) \quad \phi_o - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} x_i + v_{11}^{-1} K_{21}(w_o + \sum_{i=1}^2 \alpha_i \rho_i^{1/2} x_i)$$

$$= v_{11}^{-1} (F_2 + \omega)$$

$$=: -\frac{1}{2} \sum_{i=1}^2 \beta_i \rho_i^{-1/2} x_i + \psi_o,$$

$$(1.34) \int_{\Gamma_1} \left[ \phi_o - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} \chi_i \right] ds = B.$$

With the function spaces

$$(1.35) Z^\tau := \left\{ \{\alpha_i, \phi_o\} \mid \|\{\alpha_i, \phi_o\}\|_{Z^\tau} := \begin{cases} |\alpha_1| + |\alpha_2| + \|\phi_o\|_{H^\tau(\Gamma_1)} & \text{for } 0 \leq \tau \\ \|\phi_o - \frac{1}{2} \sum_{i=1}^2 \alpha_i \rho_i^{-1/2} \chi_i\|_{H^\tau(\Gamma_1)} & \text{for } \tau < 0 \end{cases} \right\}$$

the desired shift theorem takes the form [83, Theorem 2.5]:

Theorem 1.5: Let  $-1 < t \leq \tau + 1 < 2$ ,  $\tau \neq -1$ ,  $\tau \neq 0$ . Then the system (1.32) - (1.34) defines an isomorphism in  $H^t(\Gamma_2) \times F^\tau(\Gamma_1) \times \mathbb{R}$ .

The proof rests on Theorem 1.2 and the mapping properties of  $R_{12}$  and  $K_{2j}$ .

## §2 Improved Galerkin's Method and Galerkin Collocation with Piecewise Quadratic Functions

For Galerkin's procedure we use the finite elements (I4.7), (I4.8) with  $m = 2$  for the smooth parts  $w_o$  and  $\phi_o$  in (1.13) - (1.15). For the collision points  $z_i$  we require

$$(2.1) z_i \in \{z(j \cdot h) \mid j = 0, 1, 2, \dots, N\}$$

For convenience let us introduce the following two sets of indices:

$$(2.2) \quad I_1 := \{j \mid 0 \leq j \leq N \quad \mu_j|_{\Gamma_1} \neq 0\} ,$$

$$I_2 := \{j \mid 0 \leq j \leq N \quad \mu_j|_{\Gamma_2} \neq 0\} .$$

Now we are in the position to define the subspaces on  $\Gamma_\ell$  by

$$(2.3) \quad H_h(\Gamma_\ell) := \{\lambda_h = \sum_{j \in I_\ell} \gamma_j u_j(t) \mid_{\Gamma_\ell}\} \quad \ell = 1, 2,$$

and

$$(2.4) \quad \tilde{H}_h(\Gamma_\ell) := \{\tilde{\lambda}_h = \sum_{j \in I_\ell} \tilde{\gamma}_j u_j(t) \mid_{\Gamma_\ell} \mid \tilde{\lambda}_h(z_i) = 0, i = 1, 2\},$$

$$\ell = 1, 2 .$$

In order to formulate the modified Galerkin method for (1.13)-(1.15) we first approximate the given functions  $g_\ell, b_Y, g_{\ell h} \in H_h(\Gamma_\ell)$ ,  $\ell = 1, 2$ , requiring

$$(2.5) \quad (g_{\ell h}, u_j)_{L_2(\Gamma_\ell)} = (g_\ell, u_j)_{L_2(\Gamma_\ell)} \quad \text{for all } j \in I_\ell .$$

Then  $\tilde{g}_{\ell h} \in H_h(\Gamma_{\ell+1})$  with  $\Gamma_3 := \Gamma_1$  are chosen arbitrarily satisfying

$$\tilde{g}_{\ell h}(z_i) = g_{\ell h}(z_i), \quad i = 1, 2 ,$$

e.g. by linear functions of  $t$ .

For the smooth parts of the desired solutions  
we choose the approximations

$$(2.6) \quad w_{oh} = \sum_{j \in I_2} \tilde{\gamma}_j \mu_j(t) \text{ with } w_{oh}(z_i) = 0, i = 1, 2,$$

$$\psi_{oh} = \sum_{j \in I_1} \tilde{\beta}_j \mu_j(t) \text{ with } \phi_{oh}(z_i) = 0, i = 1, 2.$$

Now the Galerkin equations for (1.13)-(1.15) read as

$$(2.7) \quad \begin{aligned} & \int_{\Gamma_2} \left\{ w_{oh} - \frac{1}{\pi} \int_{\Gamma_2} w_{oh} d\theta_z + \frac{1}{\pi} \int_{\Gamma_1} \phi_{oh} \log|z-\zeta| ds_{\zeta} \right. \\ & + \sum_{i=1}^2 \tilde{\alpha}_i \left[ \rho_i^{1/2} \chi_i - \frac{1}{\pi} \int_{\Gamma_2} \rho_i^{1/2} \chi_i d\theta \right. \\ & \left. \left. - \frac{1}{2\pi} \int_{\Gamma_1} \rho_i^{-1/2} \chi_i \log|z-\zeta| ds_{\zeta} \right] \right\} \lambda_h ds_z \\ & = \int_{\Gamma_2} \left\{ \frac{1}{\pi} \int_{\Gamma_1} g_{1h} d\theta + \frac{1}{\pi} \int_{\Gamma_2} \tilde{g}_{1h} d\theta - \tilde{g}_{1h} \right. \\ & \left. - \frac{1}{\pi} \int_{\Gamma_1} \tilde{g}_{2h} \log|z-\zeta| ds_{\zeta} \right. \\ & \left. - \frac{1}{\pi} \int_{\Gamma_2} g_{2h} \log|z-\zeta| ds_{\zeta} - \omega \right\} \lambda_h ds_z \quad \text{for all } \lambda_h \in \tilde{H}_h(\Gamma_2), \end{aligned}$$

$$\begin{aligned}
 & \int_{\Gamma_1} \left\{ -\frac{1}{\pi} \int_{\Gamma_1} \left[ \phi_{oh} - \frac{1}{2} \sum_{i=1}^2 \tilde{\alpha}_i \rho_i^{-1/2} \chi_i(\zeta) \right] \log|z-\zeta| ds_\zeta \right. \\
 & \quad \left. + \frac{1}{\pi} \int_{\Gamma_2} w_{oh} d\theta + \sum_{i=1}^2 \tilde{\alpha}_i \frac{1}{\pi} \int_{\Gamma_2} \rho_i^{1/2} \chi_i d\theta \right\} \tilde{e}_h ds_z \\
 (2.8) \quad & = \int_{\Gamma_1} \left\{ g_{1h}(z) - \frac{1}{\pi} \int_{\Gamma_1} g_{1h} d\theta - \frac{1}{\pi} \int_{\Gamma_2} \tilde{g}_{1h} d\theta \right. \\
 & \quad \left. + \frac{1}{\pi} \int_{\Gamma_1} \tilde{g}_{2h} \log|z-\zeta| ds_\zeta \right. \\
 & \quad \left. + \frac{1}{\pi} \int_{\Gamma_2} g_{2h} \log|z-\zeta| ds_\zeta + \omega \right\} \tilde{e}_h ds_z
 \end{aligned}$$

for all  $\tilde{e}_h \in \tilde{H}_h(\Gamma_1)$  and  $\tilde{e}_h = \rho_i^{-1/2} \chi_i$ ,  $i = 1, 2$ ;

$$(2.9) \quad \int_{\Gamma_1} \left[ \phi_{oh} - \frac{1}{2} \sum_{i=1}^2 \tilde{\alpha}_i \rho_i^{-1/2} \chi_i \right] ds = - \int_{\Gamma_1} \tilde{g}_{2h} ds - \int_{\Gamma_2} g_{2h} ds = B.$$

For an asymptotic error analysis of the improved method (2.7)-(2.9) we need the approximation properties (I3.15) for  $u \in H^s(\Gamma_j)$ ,  $\mu \in H_h(\Gamma_j)$ ,  $j = 1, 2$ ,  $m = 2$  as well as for  $u \in H^s(\Gamma_j) \cap H^1(\Gamma_j)$  and  $\mu \in \tilde{H}_h(\Gamma_j)$  and also the corresponding inverse assumptions (I.3.16). In addition we need for

$$\psi_h = \psi_{oh} + \sum_{i=1}^2 \beta_i \rho_i^{-1/2} \chi_i, \quad \psi_{oh} \in \tilde{H}_h(\Gamma_1)$$

the inverse assumption

$$(2.10) \quad \|\psi_h\|_{Z^s} \leq M h^{r-s-\varepsilon} \|\psi_h\|_{Z^r}$$

with  $-2 < r \leq s < 2$  and  $\varepsilon > 0$  if  $s < 0$  and  $r \geq 0$ , otherwise

$\varepsilon = 0$ . The proof in [83, Lemma A.5] is not complete.

The complete proof can be found in [25].

With (2.10) available, a simple modification of the results in [83] yields the following error estimates [50]:

Theorem 2.1: There exists a meshwidth  $h_o > 0$  such that the Galerkin equations (2.7)-(2.9) are uniquely solvable for any  $h$ ,  $0 < h \leq h_o$ . For decreasing meshsize  $h \rightarrow 0$  we have the asymptotic error estimates

$$(2.11) \quad \begin{aligned} & \sum_{i=1}^2 |\tilde{\alpha}_i - \alpha_i| + \|\phi_{oh} - \phi_o\|_{\tilde{H}^{t-1}(\Gamma_1)} + \|v_{oh} - v_o\|_{H^t(\Gamma_2)} + |\tilde{\omega} - \omega| \\ & \leq ch^{r-t-\varepsilon} \left\{ \|g_1\|_{H^r(\Gamma_1)} + \|g_2\|_{H^{r-1}(\Gamma_2)} \right\} \end{aligned}$$

for  $1 \leq t \leq r < 2$  and any  $\varepsilon > 0$

and

$$(2.12) \quad \|\phi_{oh} - \phi_o - \sum_{i=1}^2 \frac{1}{2} (\tilde{\alpha}_i - \alpha_i) \rho_i^{-1/2} x_i\|_{H^{t-1}} + |\tilde{\omega} - \omega| \\ + \|v_{oh} - v_o + \sum_{i=1}^2 (\tilde{\alpha}_i - \alpha_i) \rho_i^{1/2} x_i\|_{H^t(\Gamma_2)} \\ \leq ch^{r-t-\varepsilon} \left\{ \|g_1\|_{H^r(\Gamma_1)} + \|g_2\|_{H^{r-1}(\Gamma_2)} \right\}$$

for  $-1 < t \leq r < 2$ ,  $t < 1$  and any  $\varepsilon > 0$  if  $1/2 < t < 1$   
and  $\varepsilon = 0$  if  $-1 < t \leq 1/2$ . The constant c is independent of  
h,  $\phi_o$ ,  $v_o$ ,  $\alpha_i$ ,  $\phi_{oh}$ ,  $v_{oh}$  and  $\tilde{\alpha}_i$  but may depend on  $\varepsilon$ .

Remark 2.1: (2.12) provides an explicit error estimate of order  $h^{1-\varepsilon}$  with any  $\varepsilon > 0$  for the stress intensity factors if  $g_1, g_2$  are given smooth enough, e.g.  $g_1 \in H^2(\Gamma_1)$ ,  $g_2 \in H^1(\Gamma_2)$ . On the other hand, the highest possible order in (2.12) is  $h^{3-\varepsilon}$ , that is two orders higher. Then  $t = -1 + \varepsilon$  and the corresponding norms on the left hand side of (2.12) are rather weak. However, the estimate (2.12) yields inner local estimates for the corresponding generated potentials (1.2) in  $\Omega$  with respect to any local norm, i.e. local super approximation of order  $h^{3-\varepsilon}$  in  $\Omega$  (see [50]). This result suggests to improve the accuracy of the stress intensity factors by an additional fitting within  $\Omega$ . Instead of fitting, the computation of the

J-integrals via our approximation of  $u$  also promises an  $h^{3-\varepsilon}$  approximation of the stress intensity factors.

For the numerical treatment of the Galerkin equations (2.7)-(2.9) one has to evaluate the entries of the stiffness matrix on the left hand sides and the weights on the right hand sides as well as of (2.5) by the use of appropriate numerical integrations. Note that in (2.7)-(2.9) on both sides appear the same double integrals if  $g_\ell, \tilde{g}_\ell$  are replaced by  $g_{\ell h}, \tilde{g}_{\ell h}$  according to (2.5).

In the following we indicate our choices of the quadrature formulas. More details can be found in [50]. We consider first the cases  $z_i \notin (\text{supp } u_j)^\circ \cup (\text{supp } u_k)^\circ$ , i.e. away from the collision points  $z_i$ .

### 2.1 The Logarithmic Standard Terms

For these terms we follow (I.4.22) and use

$$(2.13) \int_{\mathbb{R}} \int_{\mathbb{R}} \log|t-\tau| u\left(\frac{t}{h} - j\right) u\left(\frac{\tau}{h} - k\right) dt d\tau$$

$$= h^2 (\log h + w_\rho(j, k))$$

with  $\rho(j, k) = |j-k|$  the weights  $w_\rho$  in [37, Table 1] for  $m = 2$ . They are accurate up to 10 decimal digits.

## 2.2 The Regular Double Layer Weights

Because of (1.16) we integrate the corresponding weights by parts obtaining

$$(2.14) \int \mu\left(\frac{\tau}{h} - k\right) \int \mu'\left(\frac{t}{h} - j\right) \frac{d\theta}{dt} dt d\tau \\ = - \frac{1}{h} \int \mu\left(\frac{\tau}{h} - k\right) \int \mu'\left(\frac{t}{h} - j\right) \theta_{z(\tau)}(t) dt d\tau .$$

Since  $\mu$  is piecewise linear, i.e., a finite element function in the sense of [37, (2.14)] with  $m = 1$ , we use the corresponding three point integration formula with  $m = 1$ , [37, (5.15)] for the inner integral, i.e.

$$(2.15) \theta_{z(\tau)}(j) := \frac{1}{12} (\theta_z(\tilde{z}_{j+1}) - \theta_z(\tilde{z}_j))$$

$$+ \frac{5}{6} (\theta_z(\tilde{z}_{j+3}) - \theta_z(\tilde{z}_{j+2}))$$

where  $\tilde{z}_j = z(j-h)$ ,  $j = 0, 1, \dots$  and where

$$(2.16) \theta_z(\tilde{z}_{j+1}) - \theta_z(\tilde{z}_j) = \arg \left[ \frac{\tilde{z}_{j+1} - z}{\tilde{z}_j - z} \right] .$$

For the outer integration in (2.14) we use the three point integration formula [37, (5.21)] for  $m = 2$  obtaining

$$(2.17) \int u_k(z) \int u_j(\zeta) d\theta_z(\zeta) ds_z \\ \approx h \left\{ \frac{1}{8} \theta_{z_{k-1}}(j) + \frac{3}{4} \theta_{z_k}(j) + \frac{1}{8} \theta_{z_{k+1}}(j) \right\} .$$

Note that all angles in (2.17), respectively (2.15) can be evaluated explicitly via trigonometric functions.

### 2.3 Smooth Remaining Terms

The weights due to smooth remainders are of the form

$$(2.18) \int \int A(z, \zeta) u_j(\zeta) u_k(z) ds_\zeta ds_z \\ = \int \int \log \left| \frac{z(t) - z(\tau)}{t - \tau} \right| u\left(\frac{t}{h} - j\right) u\left(\frac{\tau}{h} - k\right) dt d\tau .$$

For the corresponding numerical integrations we have used four point Gaussian formulas (I4.20) although one also could apply the three point formulas with  $m = 2$  as in [37].

### 2.4 Weights Involving the Singular Elements

It remains to evaluate the weights if  $z_i$  is in the domain of integration or if the singular elements  $\rho_i^{-1/2} \chi_i$  are involved. In case of regular finite elements some of the integrals in (2.13), (2.14), (2.18) are integrated only over regions corresponding to one of

the parts  $\Gamma_1$  or  $\Gamma_2$ . In all these cases we have used four point Gaussian integration either without or with logarithmic weight function. Let us omit these details, they can be found in [50].

In case of the singular elements let us consider only one of the typical cases, for the others we again refer to [50]. Let us consider

$$(2.19) \quad I_j := \int_{\Gamma_1} \mu_j(z) \int_{\Gamma_1} \log|z-\zeta| \rho_1^{-1/2}(\zeta) \chi_1(\zeta) ds_\zeta ds_z .$$

The cut-off function  $\chi_1$  we define by a combination of a piecewise polynomial and the square root function, namely by

$$(2.20) \quad \chi_1(\zeta(t)) := \frac{1}{\sqrt{|z(t)|}} \begin{cases} 1 & \text{for } t \leq \frac{\delta}{2}, \\ \sqrt{\epsilon} \cdot v(t) & \text{for } \frac{\delta}{2} < t \leq \delta, \\ 0 & \text{otherwise,} \end{cases}$$

where  $v(t)$  is given by

$$(2.21) \quad v(t) := \frac{3}{2} \left( \frac{2t}{\delta} \right)^3 - \frac{13}{2} \left( \frac{2t}{\delta} \right)^2 + 8 \left( \frac{2t}{\delta} \right) - 2 .$$

Note that with  $\chi_1$  respectively  $v$  the whole method depends on the parameter  $\delta > 0$ , i.e. the support of  $\chi_1$ . As one of our experiments indicates,  $\delta > 0$  should be chosen not

too small in size. With (2.20), (2.21) the integral (2.19) takes the form

$$(2.22) \quad I_j = \sum_{\ell=0}^2 \int_{(j+\ell)h}^{(j+\ell+1)h} \int_0^\delta \frac{1}{\sqrt{|\dot{z}(t)|}} \frac{1}{\sqrt{t}} x_1(t) \log|z(t)-z(\tau)|$$

$$\times |\dot{z}(t)| dt \mu\left(\frac{\tau}{h} - j\right) d\tau$$

$$= \sum_{\ell=0}^2 \int_{(j+\ell)h}^{(j+\ell+1)h} \left\{ \int_0^{\delta/2} \frac{1}{\sqrt{t}} \log|z(t)-z(\tau)| dt \right.$$

$$+ \left. \int_{\delta/2}^\delta \log|z(t)-z(\tau)| v(t) dt \right\} d\tau .$$

In order to regularize the first integral in (2.22) we introduce there the new variable  $t = x^2$  arriving at

$$(2.23) \quad I_j = \sum_{\ell=0}^2 \left\{ \int_{(j+\ell)h}^{(j+\ell+1)h} 2 \int_0^{\sqrt{\delta/2}} \log|z(x^2)-z(\tau)| dx \mu\left(\frac{\tau}{h} - j\right) d\tau \right.$$

$$+ \left. \int_{(j+\ell)h}^{(j+\ell+1)h} \int_{\delta/2}^\delta v(t) \log|z(t)-z(\tau)| dt \mu\left(\frac{\tau}{h} - j\right) d\tau \right\} .$$

All outer integrations with respect to  $\tau$  in (2.23) have been executed with the regular four point Gaussian formula. For the inner integrations we have distinguished the cases  $j > 7$

and  $j < 7$ . In case  $j = 7$  we again used four point Gaussian formulas. For  $j \geq 7$  we use weighted Gaussian formulas with the logarithmic weight and 20 nodal points (see for such formulas in [73]).

## 2.5 Error Estimates for the Galerkin Collocation

In order to find the consistency estimates for our Galerkin collocation we collect all error terms corresponding to the foregoing numerical integrations.

For (2.13) let us assume that the  $w_i$  are available as accurate as required--for the presented results they are accurate up to 10 decimal digits. Thus we neglect corresponding error terms.

For the double layer weights (2.17) we can use the error estimates [37, (5.18)] with  $m = 1$  and  $m = 2$ , correspondingly, and find an error of order  $h^6$  for each weight similarly to [37, (5.20)].

For the smooth remainder terms we find an error of order  $h^8$  for each weight corresponding to the four point Gaussian formula. Analogously, the errors belonging to (2.23), i.e. to the weights involving singular elements, are of the same order  $h^8$  each.

In order to formulate the consistency estimates let us abbreviate the Galerkin equations (2.7)-(2.9) by

$$(2.24) \quad (\mathbf{A}\mathbf{V}_h, \mathbf{w}_h) = (\mathbf{f}, \mathbf{w}_h)$$

for all  $\mathbf{w}_h \in \mathbb{H}_h := \mathbb{H}_h(\gamma_2) \cup (\mathbb{H}_h(\gamma_1) \oplus \{\alpha_i \nu_i^{-1/2} \chi_i; i = 1, 2\})$

and the corresponding equations defined with the above numerical integrations by

$$(2.25) \quad (\hat{A}\hat{V}_h, W_h) = (\tilde{\phi}, W_h) .$$

Then we find as in [35, Theorem 6.2] the consistency estimate

$$(2.26) \quad |(\hat{A}U_h, W_h) - (AU_h, W_h)| \leq h \cdot \epsilon(h) \|U_h\|_{L_2} \|W_h\|_{L_2}$$

with

$$(2.27) \quad \epsilon(h) \leq c \cdot h^3 ,$$

where  $c$  denotes a constant independent of  $h$ ,  $U_h$  and  $W_h$ . This consistency in connection with the estimates (2.11), (2.12) implies the following error estimates for the solution  $\hat{\phi}_{oh}$ ,  $\hat{v}_{oh}$ ,  $\hat{\omega}$ ,  $\hat{\alpha}_i$  of the numerically integrated Galerkin equations (2.7)-(2.9).

Theorem 2.2 [50]: There exists a meshwidth  $h_0 > 0$  such that the numerically integrated Galerkin equations corresponding to (2.7)-(2.9) are uniquely solvable for any  $h$  with  $0 < h \leq h_0$ . For  $h \rightarrow 0$  we have the asymptotic error estimates

$$(2.28) \quad \sum_{i=1}^2 |\hat{\alpha}_i - \bar{\alpha}_i| + \|\phi_{oh} - \hat{\phi}_{oh}\|_{L_2(\Gamma_1)} \\ \leq c_\varepsilon h^{1-\varepsilon} \left\{ \|g_1\|_{H^2(\Gamma_1)} + \|g_2\|_{H^1(\Gamma_2)} \right\}$$

and

$$(2.29) \quad \|v_{oh} - \hat{v}_{oh}\|_{L_2(\Gamma_2)} + |\omega - \hat{\omega}| \leq ch^2 \left\{ \|g_1\|_{H^2(\Gamma_1)} + \|g_2\|_{H^1(\Gamma_2)} \right\}$$

with any  $\varepsilon > 0$ . The constants are independent of  $h$ , the data  $g_1$ ,  $g_2$  and the solutions but  $c$ , may depend on  $\varepsilon$ .

### § 3 Numerical Results

The following numerical experiments have been carried out on the IBM 370-168 computer at the Technische Hochschule Darmstadt.

For  $\Omega$  we choose the unit disc with

$$\Gamma_1 = \{z = \cos 2\pi t + i \sin 2\pi t \mid -\frac{1}{4} < t < \frac{1}{4}\}$$

$$\Gamma_2 = \{z = \cos 2\pi t + i \sin 2\pi t \mid \frac{1}{4} < t < \frac{3}{4}\}$$

$$z_1 = -i, \quad z_2 = +i.$$

Example 1:

$$(3.1) \quad U = \operatorname{Im} \sqrt{z-1} = -\rho_2^{1/2} \sin \frac{1}{2} \theta_2$$

with  $\rho_2 \cos \theta_2 = x$ ,  $\rho_2 \sin \theta_2 = 1 - y$ ,  $\rho_2^2 = x^2 + (1 - y)^2$ .

Here

$$(3.2) \quad \alpha_1 = 0, \quad \alpha_2 = -1.$$

The given data are

$$U|_{\Gamma_1} = -\rho_2^{1/2} \sin \frac{1}{2} \theta_2, \quad 0 \leq \theta_2 \leq \frac{\pi}{2}$$

and

$$\frac{\partial U}{\partial v}|_{\Gamma_2} = \frac{1}{2\rho_2^{1/2}} (x \sin \frac{1}{2} \theta_2 + y \cos \frac{1}{2} \theta_2)|_{\Gamma_2}.$$

Largest absolute errors, case  $\delta = 0.2$ :

Number of Gridpoints N+1:	40	80	160
error of $U _{\Gamma_2}$ :	$2 \cdot 10^{-2}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-4}$
error of $\phi_0 _{\Gamma_1}$ :	$10^{-2}$	$2 \cdot 10^{-3}$	$2 \cdot 10^{-4}$
error of $\alpha_i$ :	$6 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$

As expected, the errors of  $\alpha_i$  show an order  $h^{1-\varepsilon}$  of convergence.

For this example we also experimented with different  $\delta$  in case of  $N+1 = 80$  gridpoints, i.e.  $|\check{z}_{j+1} - \check{z}_j| = \frac{2\pi}{80} \approx 0.08$ .

$\epsilon$	0.2	0.15	0.1	0.075	0.05	0.01	0.001
$ \alpha_1 - \beta_1 $ :	$5 \cdot 10^{-6}$	$3 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$10^{-5}$	$10^{-5}$	$3 \cdot 10^{-4}$	$10^{-3}$
$ \alpha_2 - \beta_2 $ :	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$10^{-1}$	$10^{-3}$	$7 \cdot 10^{-1}$

The table shows increasing errors for decreasing  $\delta$ .

Correspondingly the plots of error curves show that the biggest errors are located between the boundary points corresponding to  $\delta/2$  and  $\delta$  in (2.20). These also increase with decreasing  $\delta$ .

### Example 2:

$$(3.3) \quad u = \operatorname{Re} \sqrt{2} \left\{ \frac{\sqrt{(x+iy)^2 + 1}}{x+iy+1} - \frac{1-x+iy}{1+x+iy} \right\} .$$

Here

$$(3.4) \quad \alpha_1 = \alpha_2 = 2\sqrt{2} = 2.8284 .$$

The given data are

$$u|_{\Gamma_1} = \operatorname{Re} \sqrt{2} \left\{ \frac{\sqrt{(x+iy)^2 + 1}}{x+iy+1} - \frac{1-x+iy}{1+x+iy} \right\}|_{\Gamma_1}$$

$$\text{and } \frac{\partial U}{\partial v} \Big|_{\Gamma_2} = 0 .$$

Largest absolute errors, case  $\delta = 0.2$ :

Number of grid points N+1:	40	80	160
error of $U _{\Gamma_2}$ :	$7 \cdot 10^{-2}$	$10^{-2}$	$10^{-3}$
error of $\phi_0 _{\Gamma_2}$ :	$1.5 \cdot 10^{-1}$	$7 \cdot 10^{-2}$	$4 \cdot 10^{-2}$
error of $\alpha_i$ :	$7 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-2}$

Example 3:

$$(3.5) \quad U = y^2 - x^2 .$$

For this smooth solution we have

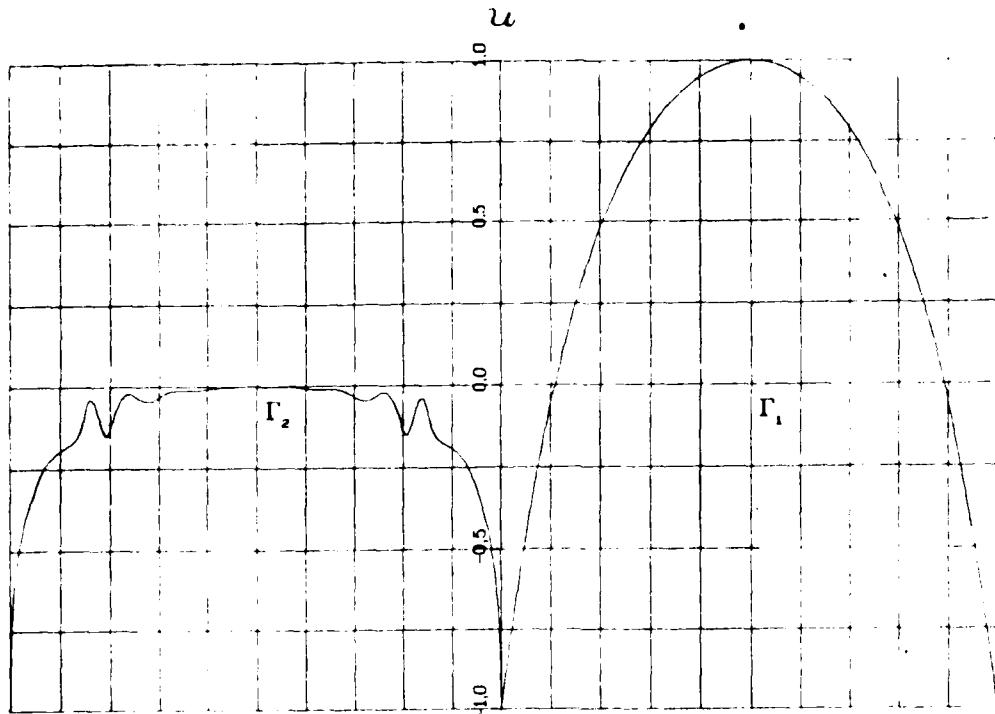
$$(3.6) \quad \alpha_1 = \alpha_2 = 0$$

and the given data are

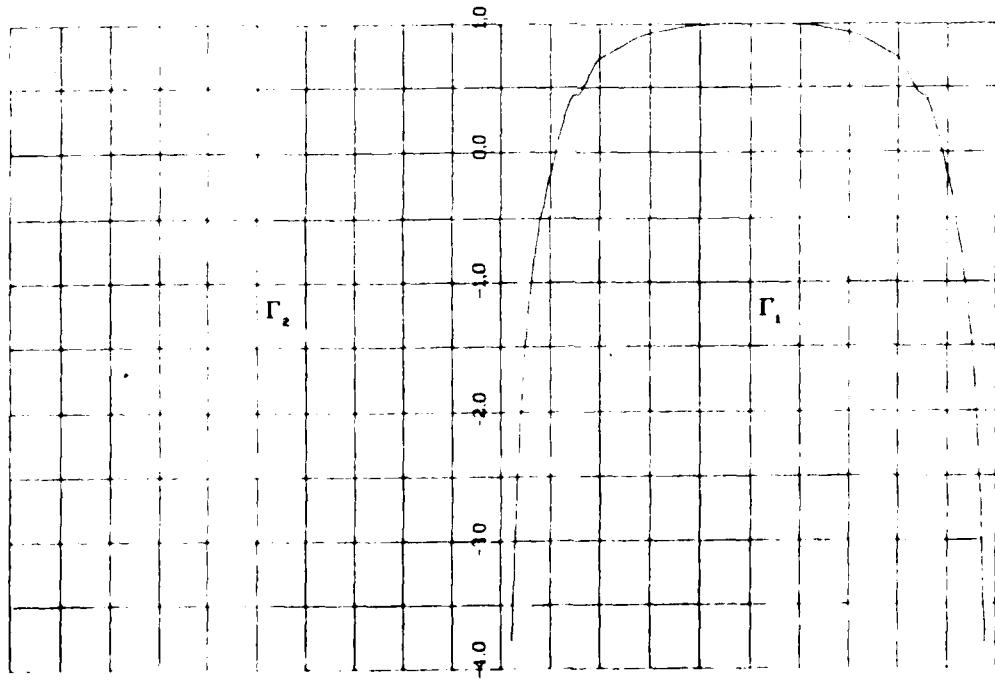
$$U|_{\Gamma_1} = y^2 - x^2|_{\Gamma_1} = \sin^2 2\pi t - \cos^2 2\pi t \quad \text{for } -\frac{1}{4} < t < \frac{1}{4} ,$$

$$\frac{\partial U}{\partial v}|_{\Gamma_2} = 2(y^2 - x^2)|_{\Gamma_2} = 2(\sin^2 2\pi t - \cos^2 2\pi t) \quad \text{for } \frac{1}{4} < t < \frac{3}{4} .$$

5.18

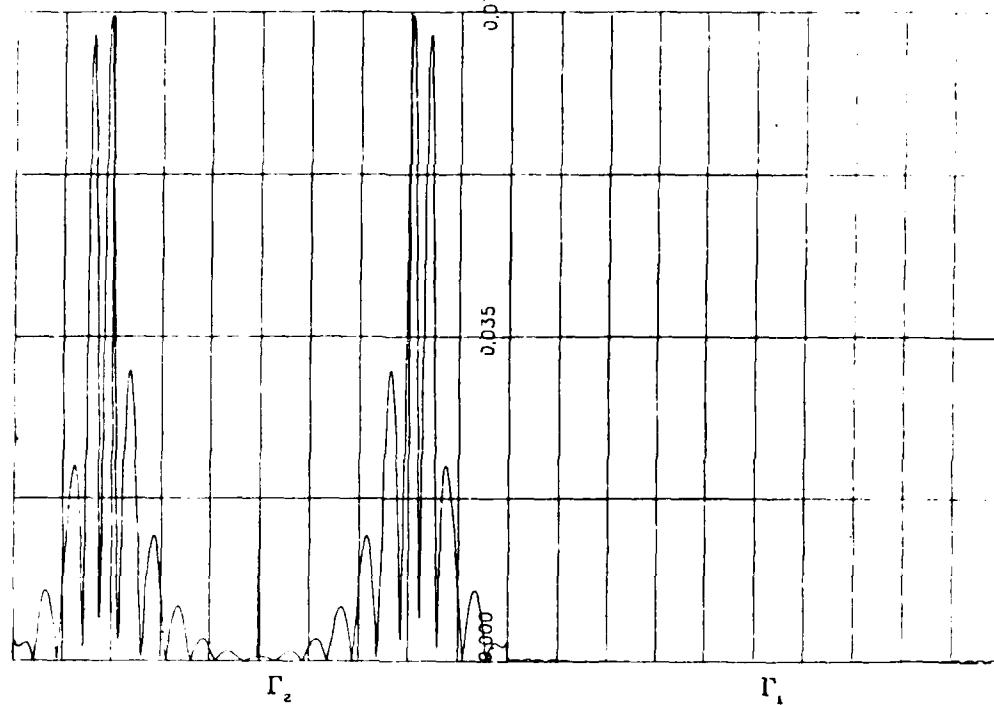


$\partial u / \partial n$

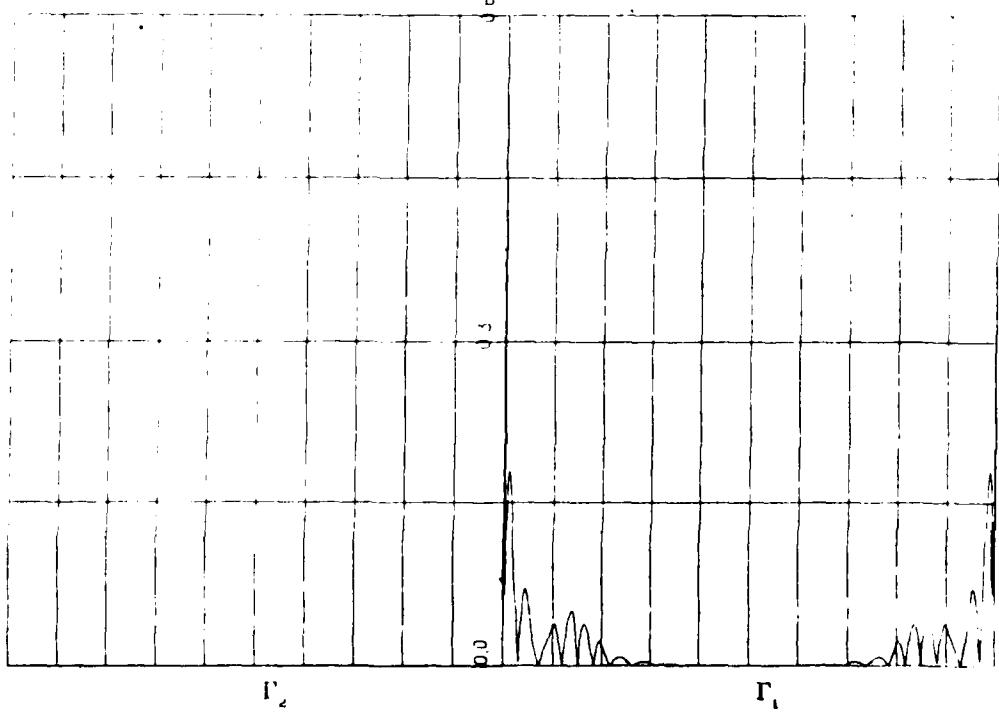


Example 2 with  $\delta = 0.2$  and 40 grid points.

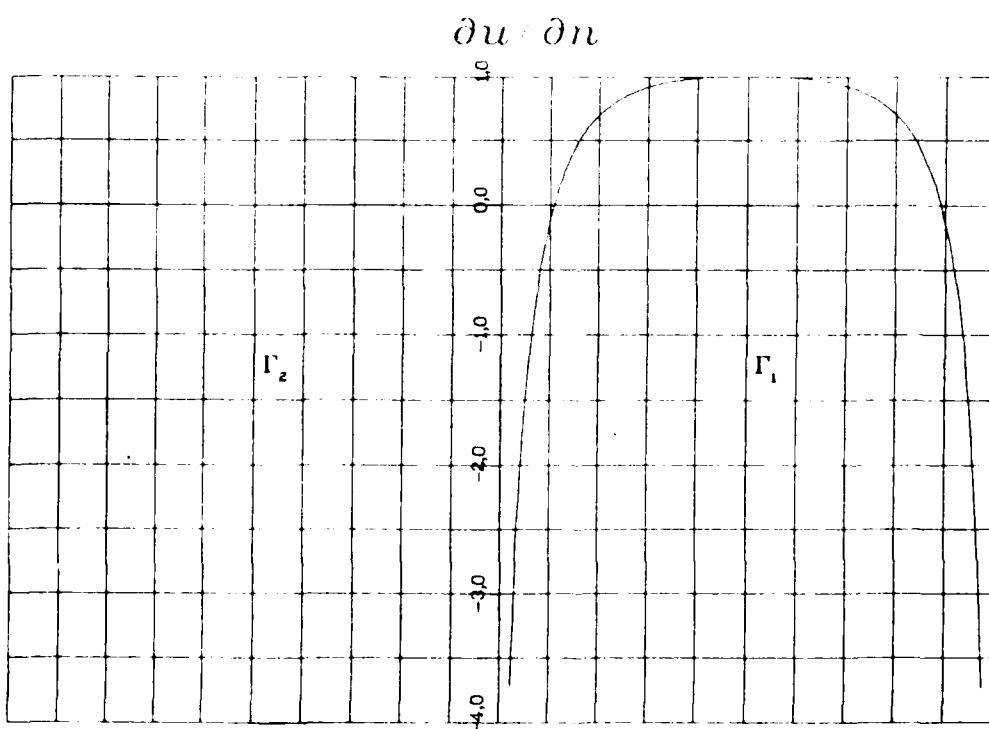
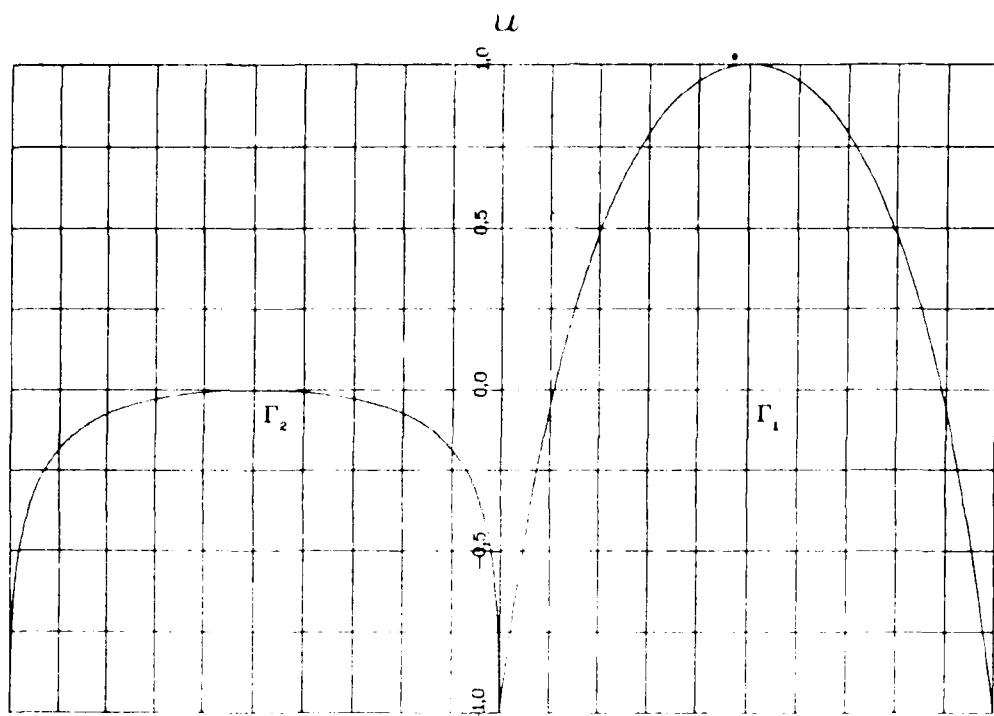
absolute error of  $u$



absolute error of  $\partial u / \partial n$

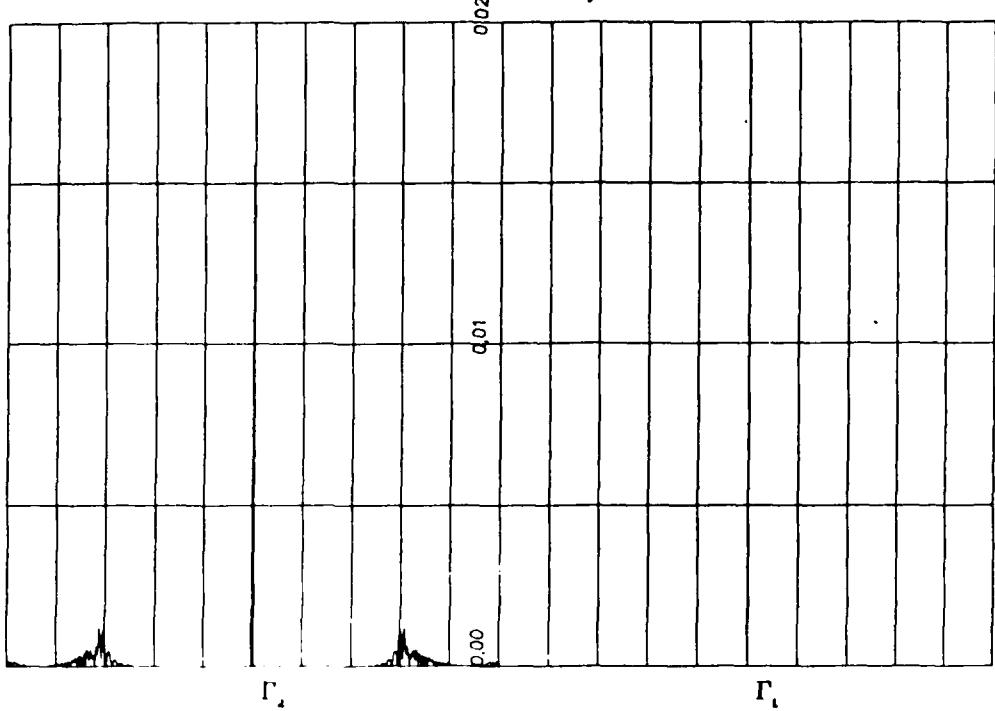


Example 2 with  $\delta = 0.2$  and 40 grid points, plots of error curves.

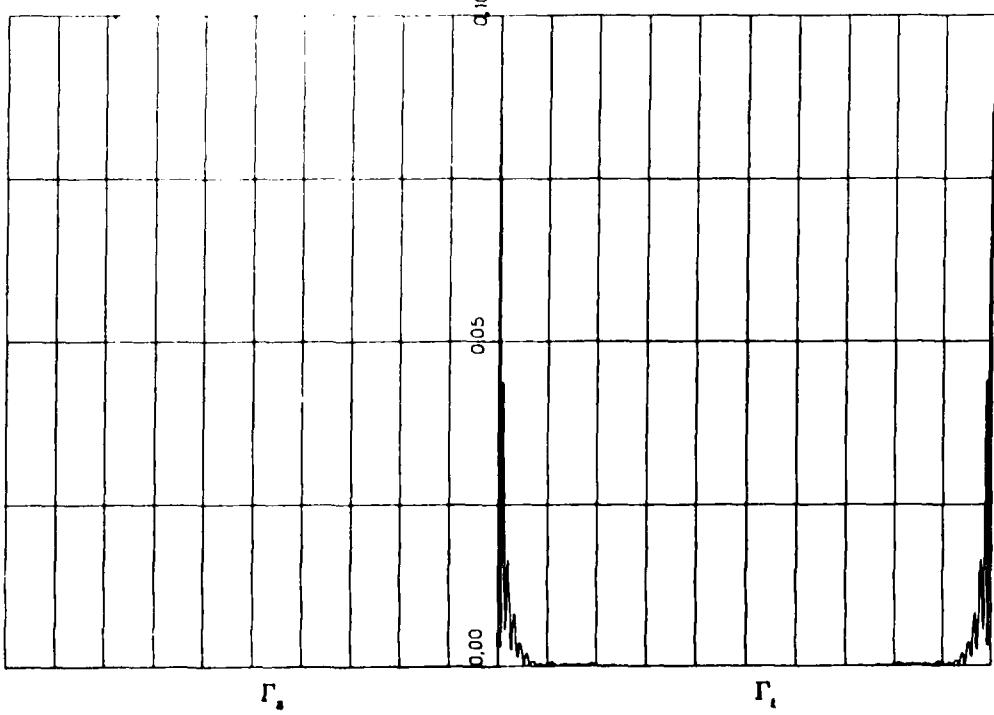


Example 2 with  $\epsilon = 0.2$  and 160 grid points.

absolute error of  $u$



absolute error of  $\partial u / \partial n$



Example 2 with  $\delta = 0.2$  and 160 grid points, plots of error curves.

Largest absolute errors, case  $\delta = 0.2$ :

Number of grid points N+1:	40	80	160
error of $U _{\Gamma_2}$ :	$2 \cdot 10^{-4}$	$3 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
error of $\phi_o _{\Gamma_1}$ :	$5 \cdot 10^{-4}$	$6 \cdot 10^{-5}$	$10^{-3}$
error of $\alpha_i, i = 1, 2$ :	$6 \cdot 10^{-4}$	$6 \cdot 10^{-6}$	$4 \cdot 10^{-4}$

In this case the errors for 160 grid points are unexpectedly too large. The reason is that the integrals (2.23) are not evaluated accurately enough. In this case the choice of  $\delta = 0.01$  improved the results significantly to  $4 \cdot 10^{-6}$  for  $U$ ,  $10^{-5}$  for  $\phi_o$  and  $2 \cdot 10^{-8}$  for  $\alpha_i, i = 1, 2$ .

§4 Plane Mixed Problems in Polygonal Domains

M. Costabel and E. Stephan extended in [24, 25] the results of [83] to Polygonal curves  $\Gamma$ .

If the smooth curve  $\Gamma$  is replaced by a polygonal, then it turns out that Lemma 1.3 is not valid anymore for (1.3), (1.4). Instead one has to eliminate  $K_{21}w_o$  from (1.14) first and then to solve the modified system

$$(4.1) \quad (I + K_{22})U|_{\Gamma_2} - V_{12} \left[ \frac{\partial U}{\partial v} \right]_{\Gamma_1} = B_1(g_1, g_2) ,$$

$$(V_{11} - K_{21}V_{12}) \left[ \frac{\partial U}{\partial v} \right]_{\Gamma_1} + K_{21}K_{22}U|_{\Gamma_2} = B_2(g_1, g_2)$$

For (4.1) they prove also Gårding's inequality corresponding to (1.27) with respect to  $L_2(\Gamma_2) \times \tilde{H}^{-1/2}(\Gamma_1)$  where  $\tilde{H}^{-1/2}(\Gamma_1)$  needs to be modified at corner points in the interior of  $\Gamma_1$  (see (4.2)). The proof is rather involved and needs in particular the Mellin transformation and a local analysis at every corner point  $z_i$ .

- 1) If  $z_i$  is a collision point of the two different boundary conditions on  $\Gamma_1$  and  $\Gamma_2$  (with or without corner) then the kernel of  $K_{22}$  vanishes identically on the adjacent straight part of  $\Gamma_2$ . Here the Mellin symbol of  $V_{11} - K_{21}V_{12}$  is positive. The set of indices for such  $z_i$  let us denote by  $I_c$ .
- 2) If  $z_i$  is an interior corner point of  $\Gamma_1$  with interior corner angle  $\omega_i$ , then let us denote by  $\Gamma_{i+}$  and  $\Gamma_{i-}$  the two straight parts of  $\Gamma_1$  adjacent to  $z_i$ . If  $\psi$  is any generalized function on  $\Gamma$  then let  $\psi_+(\rho_i)$  denote the "value" of  $\psi$  at the point on  $\Gamma_{i+}$  with distance  $\rho_i$  from  $z_i$ ;  $\psi_-$  is defined correspondingly. Let  $\mathbb{R}_+ := \{x \in \mathbb{R} | x \geq 0\}$ .

Now they define

$$(4.2) \quad \tilde{H}^{-1/2}(\Gamma^{z_i}) := \left\{ \psi \mid \psi_+ \in H^{-1/2}(\mathbb{R}_+) \wedge \psi_+ + \psi_- \in \tilde{H}^{-1/2}(\mathbb{R}_+) \right\}.$$

Then  $K_{21}V_{12}$  is compact and  $V_{11}$  is positive definite on  $\tilde{H}^{-1/2}(\Gamma^{Z_i})$ .

Using a partition of unity on  $\Gamma_1$  and pasting together  $\tilde{H}^{-1/2}(\Gamma^{Z_i})$  for all  $Z_i \in \Gamma_1$  one defines  $\tilde{H}^{-1/2}(\Gamma_1)$ .

The set of indices belonging to the interior corner points of  $\Gamma_1$  let us denote by  $I_1$ .

3) If  $Z_i$  is an interior corner point of  $\Gamma_2$  then  $K_{21}K_{22}$  becomes compact in  $L_2$  and  $\|K_{22}\|_{L_2, L_2} < 1$ , i.e.  $I + K_{22}$  becomes positive definite in  $L_2$ . The corresponding indices let us denote by  $I_2$ .

For an improvement of Galerkin's method one again expands the solution about the points  $Z_i$  and incorporates the stress intensity factors and singular functions into the integral equations (4.1) as well as into the augmented trial and test functions. Here Grisvard's representation [33] yields for the solution of (1.1) the following form:

$$\begin{aligned}
 U = & \sum_{i \in I_C} \alpha_i \rho_i^{\pi/2\omega_i} \cdot \sin(\theta_i \pi/2\omega_i) + \sum_{i \in I_C} \alpha_i \operatorname{Re}\{(z-z_i) \log(z-z_i)\} \\
 & \omega_i \neq \frac{\pi}{2}, \frac{3}{2}\pi \quad \omega_i = \frac{\pi}{2}, \frac{3}{2}\pi \\
 (4.3) \quad & + \sum_{i \in I_1} \alpha_i \rho_i^{\pi/\omega_i} \sin(\theta_i \pi/\omega_i) + \sum_{i \in I_1} \alpha_i \theta_i \\
 & \omega_i \neq \frac{\pi}{2}, \frac{3}{2}\pi \quad \omega_i = \frac{\pi}{2}, \frac{3}{2}\pi \\
 & + \sum_{i \in I_2} \alpha_i \rho_i^{\pi/\omega_i} \cos(\theta_i \pi/\omega_i) + \sum_{i \in I_2} \alpha_i \rho_i \theta_i^2 \\
 & \omega_i \neq \frac{\pi}{2}, \frac{3}{2}\pi \quad \omega_i = \frac{\pi}{2}, \frac{3}{2}\pi \\
 & + w_0
 \end{aligned}$$

where  $w_0$  is smooth. Costabel and Stephan find error estimates similar to (2.4) and (2.12). In particular for smooth enough  $g_1, g_2$  they also find convergence of order  $h^{1-\varepsilon}$  for the stress intensity factors and a maximal order  $h^{3-\varepsilon}$  with any  $\varepsilon > 0$  in appropriate weak norms.

For the details we refer to [24,25].

## §5 The Mixed Boundary Value Problem for the Three-Dimensional Laplacian

Let  $\Omega$  be a bounded simple connected domain in  $\mathbb{R}^3$  whose boundary  $\Gamma$  is a sufficiently smooth simple closed surface (at least  $C^4$ ), i.e.  $\Gamma$  is topologically equivalent to the unit sphere.  $\Gamma$  is divided into two disjoint pieces  $\Gamma_1$  and  $\Gamma_2$  such that  $\bar{\Gamma}_1 \cap \bar{\Gamma}_2 = \partial\Gamma_2 = \gamma$  defines a simple closed smooth  $C^4$  curve on  $\Gamma$ . Note that the curve  $\gamma$  now replaces the former two collision points  $z_j$ . Let us consider the classical Zaremba problem:

$$(5.1) \quad \Delta U = 0 \text{ in } \Omega, \quad U = g \text{ on } \Gamma_1 \quad \text{and} \quad \frac{\partial U}{\partial \nu} = 0 \text{ on } \Gamma_2.$$

In contrary to the two-dimensional problems, the asymptotic behaviour of  $U$  near the collision curve  $\gamma$  was not known yet. Only for half space problems with  $\Omega = \mathbb{R}^3$  one obtains the local behaviour for  $\gamma$  being a circle from the work of Sneddon and Lowengrub, see [69], for  $\gamma$  being a straight line it is given by Eskin [26]. Eskin's local asymptotic of  $U$  is

obtained via Fourier transform and Wiener-Hopf technique using distributions in weighted Sobolev spaces. For the above much more general problem Eskin's approach has been carried over by E. Stephan in [71]. Based on the formulation of the Neumann problem in [31], Baldino formulated a variational approach for the integral equations [13]. Using complex function theory Johnson investigated in [45] for the special case of a sphere an integral equation for the smooth parts of  $U$  and  $\frac{\partial U}{\partial \nu}$  on  $\Gamma$ . Based on [26], E. Stephan showed in [71] the following local behaviour of  $U$ :

Theorem 5.1: If  $g$  is smooth enough, e.g.  $g \in H^3(\Gamma)$  then the variational solution  $U \in H^1(\Omega)$  of (5.1) has the form

$$(5.2) \quad U = \alpha(s) \rho^{1/2} (\sin \frac{\theta}{2}) \chi(\rho) + v$$

with  $v \in H^{(5/2)-\epsilon}(\Omega)$  and  $\alpha \in H^{(5/2)-\epsilon}(\gamma)$  and any  $\epsilon > 0$ . Here  $s$  denotes the arc length on  $\gamma$  and  $\rho, \theta$  denote the local polar coordinates in the plane normal to  $\gamma$  and  $\Gamma$  at  $\gamma(s)$ .

Near  $\gamma$ , the transformation from  $\mathbb{R}^3$  to  $(s, \rho, \theta)$  is regular for  $\rho > 0$ .  $\chi(\rho)$  is a suitable  $C^\infty$  cut-off function with  $\chi \equiv 1$  for  $\rho$  small enough. For the further analysis we suppose without loss of generality that  $g$  is given on the whole surface  $\Gamma$ ,  $g \in H^3(\Gamma)$ . Corresponding to (5.2) the Cauchy data of  $U$  are of the form

$$(5.3) \quad U = \alpha(s)\rho^{1/2}\chi(\rho) + w_0 + g \quad \text{on } \Gamma_2$$

$$\frac{\partial U}{\partial v} = -\frac{\alpha(s)}{2}\rho^{-1/2}\chi(\rho) + \phi_0 \quad \text{on } \Gamma_1$$

with  $w_0 \in \tilde{H}^{2-\varepsilon}(\Gamma_2)$  and  $\phi_0 \in \tilde{H}^{1-\varepsilon}(\Gamma_1)$ .

Using Green's third identity and a suitable analysis of the jump condition in the frame work of Kral [47,48] and Burago, Mazja, Sapozhnikova [18] one finds the system of integral equations:

$$(5.4) \quad w_0(z) + \frac{1}{2\pi} \int_{\Gamma_2} w_0(\zeta) \left( \frac{\partial}{\partial v_\zeta} - \frac{1}{|z-\zeta|} \right) d\sigma_\zeta$$

$$+ \left\{ \alpha(s_z)\rho^{1/2}\chi(\rho) + \frac{1}{2\pi} \int_{\Gamma_2} \alpha(s_\zeta)\rho(\zeta)^{1/2}\chi \left( \frac{\partial}{\partial v_\zeta} - \frac{1}{|z-\zeta|} \right) d\sigma_\zeta \right.$$

$$+ \frac{1}{4\pi} \int_{\Gamma_1} \alpha(s_\zeta)\rho^{-1/2}\chi \frac{d\sigma_\zeta}{|z-\zeta|} - \frac{1}{2\pi} \int_{\Gamma_1} \frac{\phi_0 d\sigma_\zeta}{|z-\zeta|} \Big\}$$

$$= -g(z) - \frac{1}{2\pi} \int_{\Gamma} g(\zeta) \left( \frac{\partial}{\partial v_\zeta} - \frac{1}{|z-\zeta|} \right) d\sigma_\zeta$$

for  $z \in \Gamma_2$ ;

$$(5.5) \quad \frac{1}{2\pi} \int_{\Gamma_1} \psi_o \frac{d\sigma_\zeta}{|z-\zeta|} = \frac{1}{4\pi} \int_{\Gamma_1} \alpha_o^{-1/2} \chi \frac{d\sigma_\zeta}{|z-\zeta|}$$

$$- \frac{1}{2\pi} \int_{\Gamma_2} (\alpha_o^{-1/2} \chi + w_o) \left( \frac{\partial}{\partial v_\zeta} \frac{1}{|z-\zeta|} \right) d\sigma_\zeta$$

$$= g(z) + \frac{1}{2\pi} \int_{\Gamma} g(\zeta) \left( \frac{\partial}{\partial v_\zeta} \frac{1}{|z-\zeta|} \right) d\sigma_\zeta$$

for  $z \in \Gamma_1$ .

Since these are two integral equations with the three unknowns  $w_o$ ,  $\psi_o$  and  $\alpha$  we multiply (5.5) by  $r(z)^{-1/2}$  and integrate for fixed  $s(z)$  with respect to  $\rho$  obtaining the third equation:

$$(5.6) \quad - \int_{\bar{\rho}=0}^1 \frac{1}{4\pi} \bar{\rho}^{-1/2} \chi(\bar{\rho}) \int_{\Gamma_1} \alpha(s_\zeta) \rho(\zeta)^{-1/2} \chi(\rho(\zeta)) \frac{d\sigma_\zeta}{|z(\bar{\rho}, s) - \zeta|} d\bar{\rho}$$

$$+ \int_{\bar{\rho}=0}^1 \frac{1}{2\pi} \bar{\rho}^{-1/2} \chi(\bar{\rho}) \left\{ \int_{\Gamma_1} \frac{\psi_o d\sigma_\zeta}{|z-\zeta|} - \int_{\Gamma_2} (\alpha(s(\zeta)) \rho_\zeta^{1/2} \chi + w_o(\zeta)) \cdot \right.$$

$$\left. \cdot \frac{\partial}{\partial v_\zeta} \frac{1}{|z-\zeta|} d\sigma_\zeta \right\} d\bar{\rho}$$

$$= \int_0^1 \bar{\rho}^{-1/2} \chi \left\{ g(z) + \frac{1}{2\pi} \int_{\Gamma} g \left( \frac{\partial}{\partial v_\zeta} \frac{1}{|z-\zeta|} \right) d\sigma_\zeta \right\} d\bar{\rho}$$

on the curve  $\gamma$ .

The analysis of this system of integral equations together with appropriate finite element approximations for  $w_o$ ,  $\phi_o$  and  $\alpha$  again yields an improved boundary integral method. These will be presented in [71].

References:

- [1] Abou El-Seoud, M. S.: Numerische Behandlung von schwach singulären Integralgleichungen erster Art. Dissertation, Technische Hochschule Darmstadt, Germany 1979.
- [2] Abou El-Seoud, M.S.: Kollokationsmethode für schwach singuläre Integralgleichungen erster Art. ZAMM 59 (1979) T45-T47.
- [3] Adams, R. A.: Sobolev Spaces. Academic Press, New York, 1975.
- [4] Aleksilize, M. A.: The Solution of Boundary Value Problems with the Method of the Expansion with Respect to Nonorthonormal Functions. Nauka, Moscow (Russian) 1978.
- [5] Anselone, P. M.: Collectively Compact Operator Approximation Theory. London, Prentice Hall, 1971.
- [6] Arthur, D. W.: The solution of Fredholm integral equations using spline functions. J. Inst. Maths. Applies. 11 (1973) 121-129.
- [7] Atkinson, K. E.: A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind. Philadelphia, SIAM 1976.
- [8] Aubin, J. P.: Approximation of Elliptic Boundary-Value Problems. New York, Wiley-Interscience 1972.
- [9] Aziz, A. K. and Kellogg, R. Bruce: Finite Element Analysis of a Scattering Problem. To appear.
- [10] Babuška, I.: The finite element method with Lagrangian multipliers. Num. Math. 20 (1973) 179-192.
- [11] Babuška, I. and Aziz, A. K.: Survey lectures on the mathematical foundations of the finite element method, in "The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations" (A. K. Aziz, Ed.) 3-359, New York, Academic Press 1972.

525

- [36] Hörmander, L.: Linear Partial Differential Operators. Berlin, Springer-Verlag 1969.
- [37] Hsiao, G. C., Kopp, P. and Wendland, W. L.: A Galerkin collocation method for some integral equations of the first kind. Computing 25 (1980) 89-130.
- [38] Hsiao, G. C., Kopp, P. and Wendland, W. L.: The synthesis of the collocation and the Galerkin method applied to some integral equations of the first kind. In: C. A. Brebbia (ed.): New Developments in Boundary Element Methods. CML Publ. Southampton (1980) 122-136.
- [39] Hsiao, G. C., Kopp, P. and Wendland, W. L.: Some applications of a Galerkin-collocation method for integral equations of the first kind. In preparation.
- [40] Hsiao, G. C. and MacCamy, R. C.: Solution of boundary value problems by integral equations of the first kind. SIAM Review 15 (1973) 687-705.
- [41] Hsiao, G. C. and Wendland, W. L.: A finite element method for some integral equations of the first kind. J. Math. Anal. Appl. 58 (1977) 449-481.
- [42] Hsiao, G. C. and Wendland, W. L.: The Aubin-Nitsche lemma for integral equations. To appear in the Journal of Integral Equations.
- [43] Hsiao, G. C. and Wendland, W. L.: Super approximation for boundary integral methods. In Proc. of the Fourth IMACS Conf., 1981, to appear.
- [44] Jaswon, M. A. and Symm, G. T.: Integral Equation Methods in Potential Theory and Elastostatics. Academic Press, London 1977.
- [45] Johnson, H. L.: An integral equation formulation of a mixed boundary value problem on a sphere. SIAM J. Math. Analysis 6 (1975) 417-426.
- [46] Kohn, J. J. and Nirenberg, L.: On the algebra of pseudo-differential operators. Comm. Pure Appl. Math. 18 (1965) 269-305.
- [47] Kral, J.: The Fredholm method in potential theory. Trans. Amer. Soc. 125 (1966) 511-547.
- [48] Kral, J.: Integral Operators in Potential Theory. Springer, Lecture Notes Math. 823. Berlin, Heidelberg, New York 1980.

- [24] M. Costabel and E. Stephan: On the boundary integral method for polygonal domains. In Proc. of the Fourth IMACS Conf., 1981, to appear.
- [25] M. Costabel and E. Stephan: Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation. In preparation (Fachber. Math. THDarmstadt, Germany, Preprint No. 593, 1981).
- [26] Eskin, G. I.: Boundary Problems for Elliptic Pseudo-Differential Operators . (Russian) Nauka, Moscow 1973.
- [27] Eskin, G., Bogomilnii, A. and Zuchowizkii, S.: Numerical solution of the stamp problem . Comp. Meth. Eng. 15 (1978) 149-159.
- [28] Fichera, G.: Analisi essenziale per le soluzioni die problemi al contorno misti relativi alle equazione ed ai sistemi di equazioni del secondo ordine di tipo ellittico autoaggiuntivi. Ann. Scuola Norm. Sup. Pisa, Ser. III 1 (1949) 75-100.
- [29] Gaier, D. : Konstruktive Methoden der konformen Abbildung. Springer-Verlag, Berlin 1964.
- [30] Gaier, D.: Integralgleichungen erster Art und konforme Abbildung. Math. Zeitschr. 147 (1976) 113-129.
- [31] Giroire, J. and Nedelec, J. C. : Numerical solution of an exterior Neumann problem using a double layer potential. Math. of Comp. 32 (1978) 973-990.
- [32] Gohberg, I. C. and Feldman, I. A.: Convolution Equations and Projection Methods for their Solution. Providence, AMS Trans. 1974.
- [33] Grisvard, P.: Boundary Value Problems in Non-smooth Domains. University of Maryland, Dept. Mathematics, College Park, Md. 20742, Lecture Notes #19, 1980.
- [34] Hämmerlin, G. and Schumaker, L. L.: Procedures for kernel approximation and solution of Fredholm integral equations of the second kind, CNA Report 128, Center Numerical Analysis, Univ. Texas, Austin 1977. (Numerische Mathematik, in print).
- [35] Hildebrandt, St. and Wienholtz, E.: Constructive proofs of representation theorems in separable Hilbert space. Comm. Pure Appl. Math. 17 (1964) 369-373.

- [36] Hörmander, L.: Linear Partial Differential Operators. Berlin, Springer-Verlag 1969.
- [37] Hsiao, G. C., Kopp, P. and Wendland, W. L.: A Galerkin collocation method for some integral equations of the first kind. Computing 25 (1980) 89-130.
- [38] Hsiao, G. C., Kopp, P. and Wendland, W. L.: The synthesis of the collocation and the Galerkin method applied to some integral equations of the first kind. In: C. A. Brebbia (ed.): New Developments in Boundary Element Methods. CML Publ. Southampton (1980) 122-136.
- [39] Hsiao, G. C., Kopp, P. and Wendland, W. L.: Some applications of a Galerkin-collocation method for integral equations of the first kind. In preparation.
- [40] Hsiao, G. C. and MacCamy, R. C.: Solution of boundary value problems by integral equations of the first kind. SIAM Review 15 (1973) 687-705.
- [41] Hsiao, G. C. and Wendland, W. L.: A finite element method for some integral equations of the first kind. J. Math. Anal. Appl. 58 (1977) 449-481.
- [42] Hsiao, G. C. and Wendland, W. L.: The Aubin-Nitsche lemma for integral equations. To appear in the Journal of Integral Equations.
- [43] Hsiao, G. C. and Wendland, W. L.: Super approximation for boundary integral methods. In Proc. of the Fourth IMACS Conf., 1981, to appear.
- [44] Jaswon, M. A. and Symm, G. T.: Integral Equation Methods in Potential Theory and Elastostatics. Academic Press, London 1977.
- [45] Johnson, H. L.: An integral equation formulation of a mixed boundary value problem on a sphere. SIAM J. Math. Analysis 6 (1975) 417-426.
- [46] Kohn, J. J. and Nirenberg, L.: On the algebra of pseudo-differential operators. Comm. Pure Appl. Math. 18 (1965) 269-305.
- [47] Kral, J.: The Fredholm method in potential theory. Trans. Amer. Soc. 125 (1966) 511-547.
- [48] Kral, J.: Integral Operators in Potential Theory. Springer, Lecture Notes Math. 823. Berlin, Heidelberg, New York 1980.

[49] Lamp, U., Schleicher, T., Stephan, E. and Wendland, W. L.: The boundary integral method for a plane mixed boundary value problem. In Proc. of the Fourth IMACS Conf., R-208B, 1981, to appear.

[50] Lamp, U., Schleicher, T., Stephan, E. and Wendland, W. L.: Galerkin collocation for an improved boundary element method for a plane mixed boundary value problem, in preparation.

[51] Michlin, S. G.: Variationsmethoden der Mathematischen Physik. Berlin, Akademie-Verlag 1962.

[52] Michlin, S. G.: Approximation auf dem kubischen Gitter. Berlin, Akademie-Verlag 1976.

[53] Michlin, S. G.: On the method of least squares for multidimensional singular integral equations (Russian). In "Complex Analysis and its Applications," Izdat. Nauka, Moscow, 401-408, 1978.

[54] Michlin, S. G. and Prößdorf, S.: Singuläre Integraloperatoren. Akademie-Verlag, Berlin 1980.

[55] Nedelec, J. C.: Curved finite element methods for the solution of singular integral equations on surfaces in  $\mathbb{R}^3$ . Comp. Math. Appl. Mech. Engin. 8 (1976) 61-80.

[56] Nedelec, J. C.: Approximation des équations intégrales en mécanique et en physique. Lecture Notes, Centre de Mathématiques Appliquées. Ecole Polytechnique, Palaiseau, France 1977.

[57] Nitsche, J. A.: Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. Numer. Math. 11 (1968) 346-348.

[58] Nitsche, J. A.: Zur Konvergenz von Näherungsverfahren bezüglich verschiedener Normen. Num. Math. 15 (1970) 224-228.

[59] Noble, Ben: A Bibliography on: "Methods for solving integral equations." Math. Res. Center Tech. Report 1176 and 1177, Madison Wis. 1971.

[60] Prenter, P. M.: A collocation method for the numerical solution of integral equations, SIAM J. Numer. Anal. 10 (1973) 570-581.

- [61] Prößdorf, S.: Some Classes of Singular Equations. Amsterdam, North Holland 1978.
- [62] Prößdorf, S. and Schmidt, G.: A finite element collocation method for singular integral equations. To appear.
- [63] Prößdorf, S. and Silbermann, B.: Projektionsverfahren und die näherungsweise Lösung singulärer Gleichungen. Leipzig, Teubner 1977.
- [64] Richter, G. R.: Numerical solution of integral equations of the first kind with nonsmooth kernels. SIAM J. Numer. Anal. 17 (1978) 511-522.
- [65] Richter, G. R.: Superconvergence of piecewise polynomial Galerkin approximations for Fredholm integral equations of the second kind. Numer. Math. 31 (1978) 63-70.
- [66] Schäfer, E. : Fehlerabschätzungen für Eigenwertnäherungen nach der Ersatzkernmethode bei Integralgleichungen. Numer. Math. 32 (1979) 281-290.
- [67] Seeley, R.: Topics in Pseudo-differential Operators. In "Pseudo-Differential Operators," CIME (L. Nirenberg, ed.) Roma, Cremonese 1969.
- [68] Shaw, R., et al. (ed.): Innovative Numerical Analysis for the Engineering Sciences. The University Press of Virginia 1980.
- [69] Sneddon, I. N.: Mixed Boundary Value Problems in Potential Theory. North Holland, Amsterdam 1966.
- [70] Stakgold, I.: Boundary value Problems of Mathematical Physics. II. MacMillan Comp. New York 1968.
- [71] Stephan, E.: The boundary integral method for the three-dimensional mixed boundary value problem of the Laplacian. In preparation.
- [72] Stephan, E. and Wendland, W. L.: Remarks to Galerkin and least squares methods with finite elements for general elliptic problems. Lecture Notes Math. 564, 461-471, Berlin, Springer 1976; Manuscripta Geodactica 1 (1976) 93-123.
- [73] Stroud, A. H. and Secrest, D.: Gaussian Quadrature Formulas. Prentice-Hall, Inc. Englewood Cliffs, New York 1966.

[74] Symm, G. T.: An integral equation method in conformal mapping. *Numer. Math.* 9 (1966) 250-259.

[75] Symm, G. T.: Numerical mapping of exterior domains. *Numer. Math.* 10 (1967) 437-445.

[76] Treves, F.: *Introduction to Pseudodifferential and Fourier Integral Operators I.* Plenum Press, New York and London 1980.

[77] Vainikko, G.: On the question of convergence of Galerkin's method. *Tartu Rükl. Ül. Toim.* 177 (1965) 148-152.

[78] Voronin, V. V. and Cecch, V. A.: An interpolation method for solving an integral equation of the first kind with a logarithmic singularity. *Dokl. Akad. Nauk SSR* 216; *Soviet Math. Dokl.* 15 (1974) 949-952.

[79] Wendland, W. L.: On Galerkin collocation methods for integral equations of elliptic boundary value problems. In: J. Albrecht and L. Collatz (ed.): *Numerical Treatment of Integral Equations*. Intern. Ser. Num. Math., Birkhäuser Basel 53 (1980) 244-275.

[80] Wendland, W. L.: Asymptotic accuracy and convergence. In: C. Brebbia (ed.): *Boundary Element Methods, the State of the Art*. To appear at Pentech Press, London.

[81] Wendland, W. L.: *Elliptic Systems in the Plane*. Pitman, London, Melbourne, San Francisco 1979.

[82] Wendland, W. L. and Stephan, E.: Boundary integral method for mixed boundary value problems. In: R. Shaw et al. (ed.): *Innovative Numerical Analysis in the Engineering Sciences*. The University Press of Virginia 1980, pp. 543-554.

[83] Wendland, W. L., Stephan E., Hsiao, G. C.: On the integral equation method for the plane mixed boundary value problem of the Laplacian. *Math. Methods in the Applied Sciences* 1 (1979) 265-321.

## Defect Correction, Multigrid, and Selected Applications

by

Burton Wendroff  
Group T-7, Theoretical Division  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

### 1. Defect Correction

Defect correction is one of those deceptively simple ideas which has been around for a long time, sometimes in disguise. Many numerical algorithms use this principle, which attests its obviousness as well as its power. A definitive survey has been written by H. Stetter [20] which has aroused and renewed interest in the method of defect correction. I am going to emphasize certain formal aspects of the method, and show some applications.

The most basic defect correction algorithm is known as iterative improvement for linear systems (Forsythe and Moler [9]). Suppose that in attempting to solve the linear system  $Ax = b$  we obtain an approximation  $x_0$  which is the solution of some other system  $A_0 x_0 = b$ . For example,  $A_0$  might be the approximate LU decomposition of  $A$  obtained by Gaussian elimination. We would like to use the information contained in  $x_0$  to improve  $x_0$ . This can be done by defining a new approximation  $x_1$  by

$$(1) \quad A_0 x_1 = b + (A_0 x_0 - Ax_0)$$

A more common way to write this is to introduce the residual

$$r_0 = b - Ax_0$$

and the correction

$$d_0 = x_1 - x_0$$

so that eq. (1) becomes

$$A_0 d_0 = r_0$$

The iteration is  $r_i = b - Ax_i$ ,  $A_0 d_i = r_i$ ,  $x_{i+1} = x_i + d_i$ . However, the practical

value of this procedure is not as an iteration but as a way to reduce the error in one or two steps. Such a reduction can occur because we have the identity

$$(2) \quad A_0(x_1 - x) = (A - A_0)(x - x_0) .$$

Then for any consistent norm

$$(3) \quad \|x_1 - x\| \leq \|I - A_0^{-1}A\| \|x - x_0\| .$$

Thus, if  $A_0$  and  $x_0$  are within  $\epsilon$  of  $A$  and  $x$ , in the sense that

$$\|I - A_0^{-1}A\| < \epsilon, \|x - x_0\| < \epsilon, \text{ then}$$

$$\|x_1 - x\| \leq \epsilon^2 ,$$

and we can expect  $x_1$  to be a better approximation than  $x_0$ .

Generalizations of the identity eq. (2) are the basis of just about every successful application of defect correction.

An important step forward was taken in (Pereyra [18]), where the method is called deferred correction. Suppose that we wish to solve the differential equation

$$Lu = f$$

using a finite difference operator  $M$  to approximate the differential operator  $L$ . If

$$Mu_0 = f$$

then the analogue of (1) would be

$$\hat{Mu}_1 = f + Mu_0 - Lu_0 .$$

However,  $u_0$ , being a grid function, is not in the domain of  $L$ . What we can do is replace  $L$  by another finite difference operator  $N$  which is more accurate than  $M$ . If we define  $u_1$  by

$$(4) \quad Mu_1 = f + Mu_0 - Nu_0$$

we have instead of eq. (2) the identity

$$(5) \quad M(u_1 - u) = (M - N)(u_0 - u) + (L - N)u .$$

Suppose that for some representative grid size  $h$ , and smooth  $u$ ,

$$M_u = Lu + O(h^p) \quad \text{and} \quad Nu = Lu + O(h^q), \quad q > p .$$

Then if  $u_0 = u + O(h^p)$ , formally,

$$M(u_1 - u) = O(h^{\min(2p, q)}),$$

and if  $q \geq 2p$  we can expect  $u_1$  to be an approximation of order  $2p$ . For this to actually work there must be an error expansion of the form  $u_0 - u = h^p e$ , where  $e$  is a smooth function. In the example studied by Pereyra such an asymptotic error expansion did exist, and the indicated improvement did occur.

Note that one way to obtain a higher order operator is to set  $N = LI$ , where  $I$  is an operator defining a smooth function by interpolation from the grid function. This allows greater flexibility, and is discussed in (Frank and Ueberhuber [10]). An early application of this idea to neutron transport can be found in [16]. There, instead of increasing the order of approximation, certain poor qualitative features of  $u_0$  are improved in  $u_1$ .

The obviously attractive feature of defect correction is that with two passes through a program to solve  $Mu_0 = f$ , with different  $f$ 's, the accuracy can be increased from  $O(h^p)$  to  $O(h^{2p})$ . Apparently, only the accuracy requirement need be considered when constructing  $N$ ; stability and ease of inversion do not play a role. One question which does arise is the following: Is this the best way to achieve accuracy  $O(h^{2p})$ ? If we eliminate  $u_0$  from the equation defining  $u_1$ , we find

$$u_1 = M^{-1}(2 - NM^{-1})f .$$

Let

$$M_1^{-1} = M^{-1}(2 - NM^{-1}) .$$

Then the question is, is there an operator  $N_1$ , with the same accuracy as  $M_1$ , which in this case is  $O(h^{2p})$ , such that it is better to solve

$$(6) \quad N_1 v = f ,$$

rather than

$$(7) \quad M_1 u_1 = f ?$$

Pereyra attempts a partial answer to this very difficult question by solving a problem which was also done elsewhere by a finite element method. He correctly warns the reader not to draw too strong a conclusion from the outcome; he claims only that the comparison shows that deferred correction can be competitive. The actual problem was

$$\begin{aligned} u_{xx} + u_{yy} &= u^3 + (-2 + (1 - 2x)^2)(e^{y(1-y)} - 1 + u) \\ &\quad + (-2 + (1 - 2y)^2)(e^{x(1-x)} - 1 + u) - (e^{x(1-x)} - 1)^3(e^{y(1-y)} - 1)^3, \\ (\text{exact solution is } u(x,y) &= (e^{x(1-x)} - 1)(e^{y(1-y)} - 1)) \end{aligned}$$

on the unit square, with  $u = 0$  on the boundary. For  $M$ , Pereyra used the standard five point Laplacian, while  $N$  was a fourth order accurate difference operator. This was also solved in (Herbold [13]) using piecewise cubic finite elements to define  $N_1$ . Taking into account machine differences, eq. (7) seemed to be 100 times faster than eq. (6). The reasons that the comparison is not valid are: different iterations were used to solve the nonlinear equations; an inefficient linear system solver was used by Herbold; and the error was measured differently - at the grid points by Pereyra (apparently), and by Herbold using a much finer grid and the cubic interpolant to define intermediate points.

Before moving on to other uses of the concept of defect correction, one warning must be given. Boundary conditions and accuracy at the boundary must be given careful consideration. If this is not done the correction step will not improve the answer. An example of this can be found in (Pereyra et al. [19]).

## 2 The Multigrid Method

A very interesting and powerful application of defect correction can be found in the multigrid method for solving the differential equation

$$Lu = f$$

by means of some discretization

$$(8) \quad L_h u_h = f ,$$

defined on a grid  $G_h$ , with mesh size  $h$ . There are two parts to the idea; first, since  $u_h$  approximates  $u$  up to some truncation error, say  $O(h^p)$ , there is no point to solving eq. (8) to any better accuracy. Second, coarser grids, on which computation is relatively cheap, can be used to help with the solution of eq. (8). We will concentrate on the latter.

One starts with some relaxation procedure. For example, if  $L$  is the Laplacian and  $L_h$  is the standard five-point difference operator, then SOR might be the relaxation. After several iterations one observes that the high frequency components of the initial residual are smoothed, but then convergence slows down. The idea is to continue solving the equations on a coarser grid,  $G_{2h}$ . The crucial part is to do a defect correction on the coarse grid, that is, solve  $L_{2h} u_{2h} = f + L_{2h} J_h^{2h} u_h^0 - L_h J_h^{2h} u_h^0$ , where  $u_h^0$  is the fine grid approximation. However, the domains and ranges of these operators are wrong. So we have to choose an operator  $J_h^{2h}: u(G_h) \rightarrow u(G_{2h})$ , and then we can write

$$(9) \quad L_{2h} u_{2h} = J_h^{2h} f + (L_{2h} J_h^{2h} u_h^0 - J_h^{2h} L_h J_h^{2h} u_h^0).$$

$J_h^{2h}$  is the residual transfer operator. The grid function  $u_{2h} - J_h^{2h} u_h^0$  is the correction to be added to  $u_h^0$ ; before doing that we must define an interpolation operator  $J_{2h}^h: u(G_{2h}) \rightarrow u(G_h)$ . Then the new fine grid approximation is

$$(10) \quad v_h = u_h^0 + J_{2h}^h (u_{2h} - J_h^{2h} u_h^0) .$$

It is not necessary to obtain  $u_{2h}$  exactly, instead eq. (9) is solved by the same procedure - do several relaxation sweeps, then transfer the defect to grid  $G_{4h}$ , and so on. Only on the coarsest grid is an exact solution possibly obtained. Now we work back up through successively finer grids, using eq. (10) and additional relaxations.

Let us change notation, calling  $G_0$  the coarsest grid,  $G_1$  the next finer one, etc. The basic cycling algorithm 1) represented by the sequence  $C_N$ , where

$$C_N = G_N G_{N-1} \dots G_0 G_1 \dots G_N .$$

The full multigrid algorithm would start on the coarsest grid, as follows:  
 $C_0, C_1, C_2, \dots, C_N$ . The sequence must be terminated according to some error test.  
 Brandt uses higher order interpolation (cubic if  $L_h$  is  $O(h^2)$  accurate) each time a new fine grid is started.

How good is this? We can measure this by defining the relative efficiency as follows: Let  $\rho$  be the error reduction or spectral radius of one iteration, and let  $W$  be the work of one iteration. Define  $r = \frac{|\ln \rho|}{W}$ . The larger  $r$  the better the scheme. Suppose we measure  $W$  in units of the cost of one relaxation on the finest grid. In one (admittedly easy) example Brandt observes

$$\text{Basic cycle: } r = \frac{|\ln .25|}{8/3} = .52 .$$

In the same example the full algorithm reduces the error from .25 on the coarsest grid to .001 on the finest grid in 5.33 work units. This means

$$\text{Full algorithm: } r = 1.67 .$$

That is, in this case at least, the full algorithm is 3 times as efficient as the basic cycle. On the other hand

$$\text{SOR: } r = \frac{|\ln(1 - O(h^2))|}{1} .$$

The remarkable thing is that while  $r_{\text{SOR}} \rightarrow 0$  as  $\ln \rightarrow 0$ ,  $r_{\text{MG}}$  is asymptotically independent of  $h$ . This is proved in varying degrees of generality in [7], [12], [3], and [2].

The proper formulation of multigrid seems to be due to Fedorenko [6] and Bakhvalov [2], going back to 1961 and 1966.

Multigrid works as an acceleration of the original relaxation, and it is instructive to re-formulate it this way. I need to simplify and change the notation. First let

$$J = J_h^{2h} = \text{residual transfer}$$

$$\hat{J} = \hat{J}_{2h}^h = \text{coarse to fine interpolation}$$

$$Q = L_{2h} = \text{coarse grid operators} .$$

The relaxation sweeps are based on some splitting of  $L_h$ , say  $L_h = A - B$ . If we start with some  $w^0$  and do  $m$  relaxation sweeps, according to

$$Aw^i = Bw^{i-1} + f, \quad i = 1, \dots, m.$$

Then

$$w^m = H_1 w^0 + H_2 f$$

where

$$H_1 = (A^{-1}B)^m$$

$$H_2 = [(A^{-1}B)^{m-1} + \dots + I]A^{-1}$$

Now then, let

$$(11) \quad v^1 = H_1 v^0 + H_2 f, \quad v^0 \text{ arbitrary}.$$

On the coarse grid we have the intermediate step

$$Q\hat{u} = Jf + (QJv^1 - JL_h v^1)$$

or

$$\hat{u} = Q^{-1}Jf + Jv^1 - Q^{-1}JL_h v^1$$

$$\begin{aligned} \text{Then we set } v^1 &+ \hat{J}(\hat{u} - Jv^1) = v^1 + \hat{J}(Q^{-1}Jf - Q^{-1}JL_h v^1) \\ &= v^1 + \hat{J}Q^{-1}J(f - L_h v^1). \end{aligned}$$

This new value of  $v^1$  then starts the next sequence at relaxation sweeps.

Thus, the complete iteration is, assuming exact solution on the coarse grid, is

$$(12) \quad v^{i+1} = H_1[v^i + \hat{J}Q^{-1}J(f - L_h v^i)] + H_2 f.$$

Note that this is consistent: if  $L_h v = f$  then  $v = H_1 v + H_2 f \Rightarrow L_h v = f$ . This also shows the absolute necessity of transferring the defect to the coarse grid. Let

$$C = H_1[I - \hat{J}Q^{-1}JL_h].$$

Recall that the efficiency is  $r = \frac{|\lambda_{\max}(C)|}{w}$ , where  $\rho(C)$  = spectral radius of  $C$ . It has been proved in varying degrees of generality that there exists  $\bar{\rho}$  independent of  $h$  such that  $\rho(C) < \bar{\rho} < 1$ , even for the completely recursive algorithm. A very neat heuristic estimate of  $r$  has been given by Brandt [3], as follows: Let  $\mu$  be the smoothing factor of one relaxation sweep. Then after all the grids have been visited all the frequency components have been reduced by  $\mu^m$ , i.e.  $\rho(C) = \mu^m$ . In two dimensions the work, relative to the work of one relaxation sweep is

$$(13) \quad w = m \left(1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \dots\right) \approx \frac{4m}{3} .$$

So  $r = \frac{|\lambda_{\max}|}{4m/3}$ . This has proved to be very reliable in practice.

### 3. Higher Order and Multigrid

Brandt has also shown how one can combine multigrid with the use of defect correction to get higher order accuracy. If  $N_h$  is the higher order difference operator then the coarse grid difference equation becomes

$$L_{2h} u_{2h} = J_h^{2h} f + [L_{2h} J_h^{2h} u_h^o - J_h^{2h} N_h u_h^o] .$$

Brandt uses

$$J_h^{2h} N_h = \frac{4}{3} J_h^{2h} L_h - \frac{1}{3} L_{2h} J_h^{2h} .$$

This is formally fourth order accurate if  $L_h$  is 2<sup>nd</sup> order accurate. This is called tau-extrapolations. The coarse grid equation becomes

$$(14) \quad L_{2h} u_{2h} = J_h^{2h} f + \frac{4}{3} [L_{2h} J_h^{2h} u_h^o - J_h^{2h} L_h u_h^o].$$

Here is a sample computation. The test problem is poisson's equation on a rectangle with Dirichlet data.  $L_h$  is the standard five-point operator,  $J$  is injection, and  $\hat{J}$  is linear interpolation except when beginning a new fine grid, at which point cubic interpolation is used (with or without tau-extrapolation). The only change in strategy is to use eq. (14) instead of

eq. (9) the first time the fine grid residual is transferred to the next coarse grid. This is schematized below and the errors are shown.

Grid		Error Usual	Error Tau	Ratio
0				
1		e	e	1
<u>Start 2 with cubic interpolation</u>				
from 1				
2 relax				
(15)	—————→			
1 (9)	—————→			
0				
1		.32e	.2e	1.6
2				
<u>Start 3 with cubic interpolation</u>				
from 2				
3 relax				
(15)	—————→			
(9)	—————→			
1				
0				
1				
2				
3	—————	.72e	.02e	3.6
4				
.				
.				
.				
4	—————	.017e	.0019e	8.9
5				
.				
.				
.				
5	—————	.004e	.0002e	20

Although the accuracy is considerably enhanced by tau-extrapolation it is not fourth order, since the latter would produce ratios increasing by factors of four. The reason for the loss of accuracy is that eq. (14) is not solved exactly, that is, instead of inverting  $L_{2h}$  the multigrid algorithm inverts some approximation  $\tilde{L}_{2h}$ . Thus,

$$\begin{aligned}\tilde{L}_{2h}^{-1}L_h - Ju &= (L_{2h}J - JN_h)(u_h^2 - u) + (JL - JN_h)u \\ &\quad + (L_{2h} - \tilde{L}_{2h})Ju .\end{aligned}$$

The last term can be reduced to  $O(h^{2p})$  only by increasing the number of multigrid cycles.

What is needed here is a good test of, for example, the incomplete tau-extrapolation just described, the more accurate tau-extrapolation, and solving  $N_h u_h = f$  as efficiently as possible.

#### 4. The Coarse Grid Operator

Nicolaides [17] and Hackbusch [11] have observed that if instead of using  $Q = L_{2h}$ , we set

$$(15) \quad Q = JL_h^3$$

then  $I - \hat{J}Q^{-1}JL_h$  annihilates the range of  $\hat{J}$ . In addition, the residual of the corrected solution vanishes when transferred to the coarse grid, that is,

$$J[L_h(u^i + \hat{J}Q^{-1}J(f - L_h u^i)) - f] = 0 .$$

Alcouffe et al. [1] found that (15) was necessary in order to obtain the predicted convergence rate in a problem in which the coefficients were discontinuous and jumped by orders of magnitude.

The mappings  $J$  and  $\hat{J}$  and the relaxation splitting must be properly chosen for all this to work. This still seems to be an art, as can be seen in some of the applications to physical problems described in [1], and [5], which will be presented later. Here is a simple example. Take

$$(L_h u)_i = h^{-2} (u_{i+1} - 2u_i + u_{i-1}) .$$

Let the even points be the coarse grid, and let  $\hat{J}$  be linear interpolation.  
Thus

$$(\hat{J}u)_i = \begin{cases} u_i, & i \text{ even} \\ \frac{1}{2}(u_{i+1} + u_{i-1}), & i \text{ odd} \end{cases} .$$

Let the residual weighting operator  $J$  be defined by

$$(Ju)_i = \frac{1}{4} u_{i-1} + \frac{1}{2} u_i + \frac{1}{4} u_{i+1} , \quad i \text{ even} .$$

Then it is easy to see that

$$L_{2h} = JL_h\hat{J}$$

and

$$(15a) \quad J = \hat{J}^* = (\frac{1}{2}\hat{J})^T ,$$

$$\text{if } (a,b)_h = \sum a_i b_i h .$$

The reader might turn to section 6 to see how the averaging in eq. (15) arises naturally in a problem with variable coefficients. Eqs. (15) and (15a) are also theoretically useful, as in [12].

##### 5. As Application to Fluid Dynamics

The multigrid method has the ability, in principle, to take an existing finite difference code in which relaxation iterations use a large fraction of the running time, and speed it up considerably without making a major revision of the code. To see if this were really true in practice, my colleagues Joel Dendy and Hans Ruppel, together with Achi Brandt, incorporated the multigrid algorithm into the SOLA code.

Some of the results of this work, reported in [1], are given here, together with some additional information given me by Joel Dendy. SOLA solves the incompressible Navier-Stokes equations, which are

$$u_x + v_y = 0$$

$$u_t + (u^2 + p)_x + (uv)_y = g_x + v[u_{xx} + u_{yy}]$$

$$v_t + (v^2 + p)_y + (uv)_x = g_y + v[v_{xx} + v_{yy}]$$

$$v = 0, \frac{\partial v}{\partial y} = 0$$

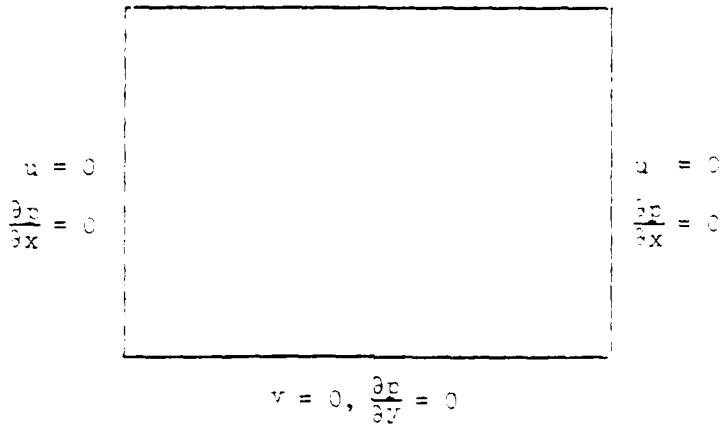


Figure 1

with boundary conditions shown in Fig. 1. The difference equations are semi-implicit, as follows:

$$(16) \quad \frac{1}{\Delta x} \left( u_{i,j}^{n+1} - u_{i-1,j}^{n+1} \right) + \frac{1}{\Delta y} \left( v_{i,j}^{n+1} - v_{i,j-1}^{n+1} \right) = 0$$

$$(17) \quad u_{i,j}^{n+1} + \frac{\Delta t}{\Delta x} \left( p_{i+1,j}^{n+1} - p_{i,j}^{n+1} \right) = a_{i,j}^n$$

$$(18) \quad v_{i,j}^{n+1} + \frac{\Delta t}{\Delta y} \left( p_{i,j+1}^{n+1} - p_{i,j}^{n+1} \right) = b_{i,j}^n .$$

The quantities  $a_{i,j}^n$  and  $b_{i,j}^n$  contain all the information from the previous time step; their exact form is irrelevant to this discussion. The grid structure is shown in Figure 2.

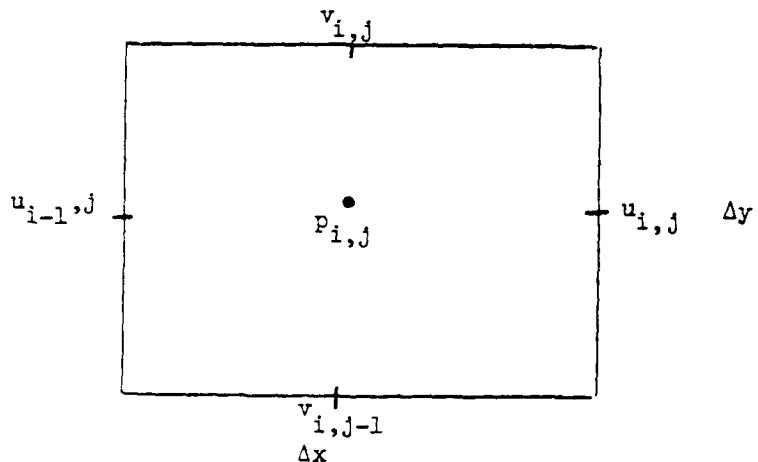


Figure 2

Note that by using eq. (17) and eq. (18) to eliminate the velocities  $u^{n+1}$  and  $v^{n+1}$  from eq. (16) we have

$$(19) \quad L_h p = c$$

where  $L_h$  is the five point Laplacian. SOLA solves this by an iteration on  $p, u$ , and  $v$  which is equivalent to successive over-relaxation for eq. (19). This iterative procedure was maintained in the multigrid implementation, the only change being that a residual appears in eq. (16) on the coarse grids. To keep the proper relationship between velocities and pressure, no residuals are introduced in eq. (17) or eq. (18), and these equations are used to define the corrected fine grid velocities once the pressure has been corrected. The grid structure and the relation between coarse and fine grids is shown in Figure 3.

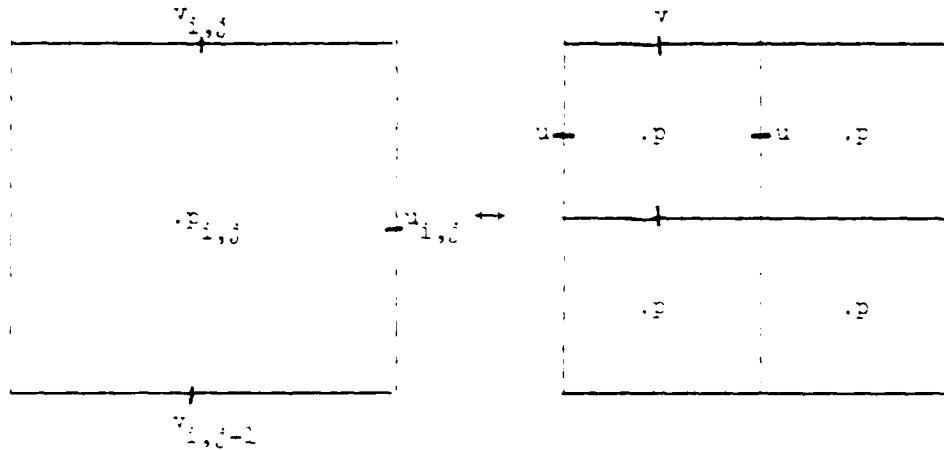


Figure 3

The residual of eq. (16) is located at the same grid points as  $p$ . The residual weighting (the interpolation  $J$ ) is obtained by defining the coarse grid residual as the equally weighted average of the four neighboring fine grid residuals except at the boundary where a special weighting was used which is discussed below. The operator  $\hat{J}$  (operating only on the pressures) was defined by bilinear interpolation. The efficiency predicted by eq. (13) was achieved. The procedure was non-adaptive, that is, the iteration was started on the finest grid.

The boundary conditions on  $u$ ,  $v$ , and  $p$  require that  $a_{i,j}$  and  $b_{i,j}$  vanish at the sides and top respectively. This insures that the sum over the grid of  $c_{i,j}$  is zero, which is necessary for (19) to have a solution. Because of the unequal residual weighting this consistency condition is not satisfied on the coarse grids, except in the limit. This causes no problem since exact solutions are not sought on the coarse grids.

Equal weights at the boundaries caused a 14% loss in the efficiency predicted by eq. (13). The success of the unequal weights used brings up some interesting points, although we cannot provide a clean argument for that success. The actual weights are shown in Figure 4.

4/9	1/3	1/3	1/3
1/3	1/4	1/4	1/4
1/3	1/4	1/4	1/4
1/3	1/4	1/4	1/4

Figure 4

Note that the weights in the coarse cells at the boundary do not add to one.

Let  $j_0$  be the local sum of the residual weights; in Figure 3,  $j_0$  is 1 at interior coarse cells, 1.35 at corners, and 1.17 at the edges. In [3,5] a heuristic argument is given to show that the effect of the iteration matrix  $C$  of (12a) is approximately  $|1 - j_0|$  when applied to the smoothest grid functions, therefore if  $j_0 = 1$ ,  $C$  will surely reduce the smoothest part of the error. It seems that all that is really necessary is that  $|1 - j_0|$  not be too large.

The weights are a bit mysterious, but they can be obtained by a more or less convincing argument which we present for the one-dimensional case. Suppose the differential equation is  $p_{xx} = f$ , with  $p_x = 0$  at the boundaries.

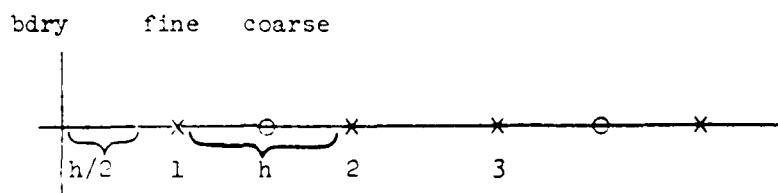


Figure 5

Referring to Figure 5, let

$$(L_h p)_i = \frac{p_{i+1} - 2p_i + p_{i-1}}{h^2}, \quad i = 2, 3, \dots,$$

and

$$(L_h p)_1 = \frac{p_2 - p_1}{h^2}.$$

Now,

$$\frac{1}{h} \int_{x_{i-h}}^{x_{i+h}} p_{xx} dx = \frac{(p_x)_{i+1} - (p_x)_{i-1}}{2h} \approx \frac{(p_{i+1} - p_i) - (p_i - (p_{i+1}))}{2h^2}, \quad i = 2.$$

and

$$\frac{1}{\frac{3h}{2}} \int_{x_{i-h}-\frac{h}{2}}^{x_{i+\frac{1}{2}}} p_{xx} dx = \frac{(p_x)_{i+\frac{1}{2}} - (p_x)_{i-\frac{1}{2}}}{3h/2} \approx \frac{p_{i+\frac{1}{2}} - p_{i-\frac{1}{2}}}{\frac{3h}{2}}.$$

This means that if  $(L_h p)_2$  is given weight 1 then  $(L_h p)_1$  should have weight  $\frac{1}{3}$ . The analogous argument in two dimensions gives the indicated weights.

The authors of [5] also implemented multigrid into the SOLA-ICE code, which is a compressible flow version of SOLA. This was not straightforward. Difficulties were encountered on the coarsest grid which could not be overcome by doing a direct solution because of the peculiar nature of the SOLA-ICE algorithm, the latter having been dictated by a desire to maintain an iteration similar to SOLA. The authors finally hit upon a technique of shifting  $p$  which improved both the original algorithm and the multigrid version to the point that the correct convergence rate was obtained.

We should point out that it is possible to take a more natural approach (from the point of view of a numerical analyst) to the solution of semi-implicit difference schemes. The nonlinear difference equations can be solved by Newton's method, however, it is important to make the right choice of variables about which to linearize. For many problems  $p = \epsilon^{-1}(p-p_0)$ ,  $\epsilon \ll 1$ .

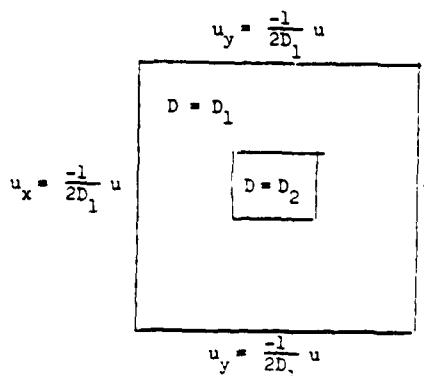
In this case linearization around  $\rho$  produces an ill-conditioned Jacobian, so that one should linearize around  $p$ . This procedure is followed in [15] where a difficult two-phase flow problem is solved. Since the method also involves relaxation oscillations it should be possible to apply the multigrid concept there also.

#### 6. Neutron Diffusion

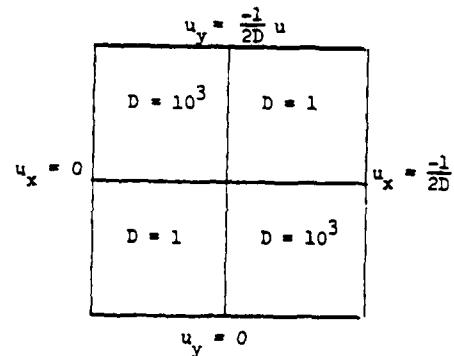
A difficult neutron diffusion problem was done successfully by the multigrid method in [1]. The problem is

$$-\nabla \cdot (D\nabla u) + \sigma u = f .$$

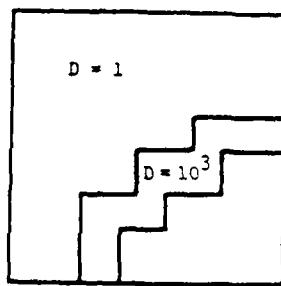
Some sample configurations and boundary conditions are shown in Figure 6.



(a)  $D_2/D_1$  ranged from  $10^{-4}$  to  $10^4$



(b) Four Corners



(c) Staircase

Figure 6

Because of the large jumps in the coefficient D there is no easy way to define the coarser grid operators, therefore the authors used eq. (15). This, together with eq. (15a), at least reduces the variability of the problem to the choice of  $\hat{J}$ . It was observed that with  $\hat{J}$  taken to be the bilinear interpolation operator the multigrid iteration either failed to accelerate the lexicographic SOR iteration, or even failed to converge at all. We can gain some insight into this problem by considering the one-dimensional problem

$$\frac{d}{dx} \left( D \frac{du}{dx} \right) = f$$

where D is a step function with jumps at the fine grid points, as in Figure 7.

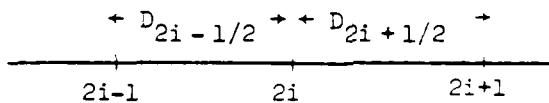


Figure 7

The fine grid difference operator (away from the boundaries) is defined by

$$\left( h^2 L_h u \right)_i = D_{i+1/2} (u_{i+1} - u_i) - D_{i-1/2} (u_i - u_{i-1}) .$$

Let the coarse grid consist of the even-indexed grid points. Consider the following method for solving

$$\left( L_h u \right)_i = f_i .$$

First, choose anything for the even indices, say  $u_{2i}^o$ , and define  $u_{2i+1}^o$  by relaxation; thus,

$$\left( L_h u^o \right)_{2i} = f_{2i} + r_i$$

$$\left( L_h u^o \right)_{2i+1} = f_{2i+1}$$

Next, eliminate the odd variables. If we define  $L_{2h}$  by

$$(L_{2h}u)_i = \frac{D_{i-1/2}}{D_{i-3/2} + D_{i-1/2}} (L_h u)_{i-1} + (L_h u)_i + \frac{D_{i+1/2}}{D_{i+1/2} + D_{i+3/2}} (L_h u)_{i+1}$$

for  $i$  even only, then

$$(20) \quad (L_{2h}u^o)_i = r_i + \left[ \frac{D_{i-1/2}}{D_{i-3/2} + D_{i-1/2}} f_{i-1} + f_i + \frac{D_{i+1/2}}{D_{i+1/2} + D_{i+3/2}} f_{i+1} \right]$$

and the left side involves only even indices. Now, solve exactly the coarse grid correction equations

$$(L_{2h}v)_i = -r_i, \quad i \text{ even}.$$

Then if

$$u_i = (u^o + v)_i, \quad i \text{ even}$$

and if  $u_i$ ,  $i$  odd, is defined by relaxation, we will have obtained the exact solution.

The following is easily verified. Let the bracketed terms in eq. (20) define the residual transfer  $J$ . The relaxation at the odd points defines an interpolation operator  $\hat{J}$ , and  $J = \hat{J}^T$ , and  $L_{2h} = JL_h\hat{J}$ .

This can be summarized by the statement that the appropriate choice of relaxation strategy and interpolation  $J$  produces the exact solution in one iteration, assuming exact solution on the coarse grid. Since the coarse grid can be treated in the same way, exact solution can be obtained in one full cycle. Furthermore,

$$(h^2 L_{2h}u)_i = \frac{D_{i+1/2}D_{i+3/2}}{D_{i+1/2} + D_{i+3/2}} (u_{i+2} - u_i) - \frac{D_{i-1/2}D_{i-3/2}}{D_{i-1/2} + D_{i-3/2}} (u_i - u_{i-2}).$$

Apart from a missing factor of 2, the coefficients are the harmonic averages of the  $D$ 's, which are well-known to be the precisely correct averages to use. While none of this carries over to two dimensions, it would seem reasonable to try to stay close to this formulation without constructing an algorithm that is too complicated. The method arrived at in [1] does just that.

Consider Figure 8. Suppose the coarse grid points A,B,C,D, have been found. A feasible procedure would be to define the interpolant at 1,2,3,4,5, by relaxation of the fine grid difference operator centered at each point. Instead, the authors chose to lump the operator centered at 1 into a 3-point operator involving A,1, and D and then used that to define the interpolant at 1. That is, if the operator becomes  $au_A - bu_1 + cu_D$ , then they set  $u_1 = (au_A + cu_D)/b$ . The corresponding interpolations are done at 2,3, and 4. The full difference operator centered at 5 is then used to define  $u_5$ . With this definition of  $\hat{J}$  the authors then took  $J = \hat{J}^T$ , and  $L_c = \hat{J}^T L_f J$ , where  $L_f$  is the fine grid operator which will in general itself have been defined in this way from still finer grids.

The computational results are quite impressive. For some of the fairly hard problems the error reduction (spectral radius) of one cycle is .1 with an efficiency matching the efficiency of the standard multigrid algorithm for the constant coefficient Laplacian.

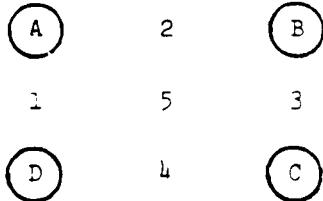


Figure 8

The method of interpolating from the coarse to fine grid described above recognizes that an elliptic difference operator defines a natural interpolation. Elliptic interpolation has been shown by J. M. Hyman [14] to be a practical way to advance from coarse to fine grids when solving Poisson's equation. A very efficient checkered relaxation form of multigrid, presented in [16] also exploits elliptic interpolation. However, this would not work here because we are not dealing with the five-point Laplacian.

7      Acknowledgment

I would like to thank Joel Dendy, Jr., for patiently answering my many questions about the multigrid method. I am also grateful to Mac Hyman and Blair Swartz for discussions about and comments on the manuscript.

## References

1. Alcouffe, R. E., Achi Brandt, J. E. Dendy, Jr., and J. W. Painter, The multigrid methods for the diffusion equation with strongly discontinuous coefficients, SIAM Journal on Scientific and Statistical Computing, to appear.
2. Bakhvalov, N. S., On the convergence of a relaxation method with natural constraints on the elliptic operator, Zh. vych. mat. 6, 1966, pp. 861-865; English translation, USSR Computational Mathematics and Mathematical Physics 6, 1966, pp. 131-135.
3. Brandt, A., Multi-level adaptive solutions to boundary-value problems. Math. Comp. 31, 1977, pp. 333-390.
4. Brandt, A., Multi-level adaptive techniques (MLAT) and singular perturbation problems, Proceedings of the Conference on Numerical Solution of Singular Perturbation Problems, University of Nijmegen, The Netherlands, 1978.
5. Brandt, A., J. E. Dendy, Jr., and H. Ruppel, The multigrid method for semi-implicit hydrodynamics codes, J. Comp. Phys. 34, 1980, pp. 348-370.
6. Fedorenko, R. P., A relaxation method for solving elliptic difference equations, Zh. vych. mat. 1, 1961, pp. 922-927; English translation, USSR Computational Mathematics and Mathematical Physics 1, 1962, pp. 1092-1096.
7. Fedorenko, R. P., The speed of convergence of one iterative process, Zh. vych. mat., 4, 1964, pp. 559-564; English translation, USSR Computational Mathematics and Mathematical Physics 4, 1964, pp. 227-235.
8. Foerster, F., K. Stuben, and U. Rüthenberg, Non-Standard multigrid techniques using checkered relaxation and intermediate grids, to appear in "Elliptic Problem Solvers", Martin Schultz, ed., Academic Press, New York 1980.
9. Forsythe, G., and C. B. Moler, "Computer Solution of Linear Algebraic Systems", Prentice Hall, 1967.
10. Frank, R. and C. W. Ueberhuber, Collocation and iterated defect correction, In: Numerical Treatment of Differential Equations, edited by R. Bulirsch, R. D. Grigorieff, J. Schroder, pp. 19-34. Lecture Notes in Mathematics, Vol. 631, Springer 1978.
11. Hackbusch, W., On the multi-grid method applied to difference equations, Computing 20, 1978, pp. 291-306.
12. Hackbusch, W., Convergence of multi-grid iterations applied to difference equations, Math. Comp. 34, 1980, pp. 425-440.

13. Herbold, R. J., Consistent Quadrature Schemes for the Numerical Solution of Boundary Value Problems by Variational Techniques, Ph.D. Thesis, Case Western Reserve University, Cleveland, Ohio, 1968.
14. Hyman, J. M., Mesh refinement and local inversion of elliptic partial differential equations, J. Comp. Phys. 23 (1977), pp. 124-134.
15. Liles, D. R., and Wm. H. Reed, A semi-implicit method for two-phase fluid dynamics, J. Comp. Phys. 26, 1978, pp. 390-407.
16. Miller, W. F., Jr., and Wm. H. Reed, Ray-Effect mitigation methods for two-dimensional neutron transport theory, Nucl. Sci. Eng., 62 (1977), pp. 391-411.
17. Nicolaides, R. A., On multiple grid and related techniques for solving discrete elliptic systems, J. Comp. Phys. 19, 1975, pp. 418-431.
18. Pereyra, V., Highly accurate numerical solution of casilinear elliptic boundary-value problems in n dimensions, Math. Comp. 24, 1970, pp. 771-783.
19. Pereyra, V., W. Proskurowski, O. Widlund, High order fast Laplace solvers for the Dirichlet problem on general regions, Math. Comp. 31, 1977, pp. 1-16.
20. Stetter, H. J., The defect correction principle and discretization methods, Numer. Math. 29, 1978, pp. 425-443.

Galerkin-Finite Element Solution  
of Nonlinear Evolution Problems

Miloš Zlámal

Computing Center of the Technical  
University in Brno, Czechoslovakia

## I. Introduction

The generalized Galerkin method for the solution of evolution problems consists of the following steps: 1) We formulate the given problem in a variational form. 2) We discretize the problem in space, i.e. we consider a family  $\{V^h\}$ ,  $0 < h < h^*$ , of finite dimensional subspaces of the basic Banach space  $V$  such that  $\lim_{h \rightarrow 0_+} \text{dist}(V^h, V) = 0$   $\forall v \in V$  and in  $V^h$  we define a semidiscrete solution by means of a discrete analog of the variational formulation determining the exact solution. 3) To compute this solution means to solve a system of ordinary differential equations. Solving this system numerically we get a completely discretized approximate solution. In case of nonlinear problems the application of linear multistep methods has advantage in that we are often able to linearize the resulting scheme without lowering the accuracy. We restrict ourselves to a narrow class of linear multistep methods: to A-stable methods. These methods lead to unconditionally stable schemes fulfilling certain energy inequalities. Both these properties are desirable, the other providing a simple way for the derivation of apriori error estimates.

First we describe a class of linear multistep methods considered in the sequel. Then we deal with the nonlinear heat equation and with the time-dependent Navier-Stokes equations. In both cases linearization is possible without lowering the accuracy. Afterward, we investigate more details the solution of quasistationary nonlinear magnetic fields. Finally, we mention some results concerning a nonlinear hyperbolic equation.

## II. A-stable Linear Multistep Methods

The characteristic polynomials of one-step consistent methods are

$$(1) \quad \varrho(\xi) = \xi - 1, \quad \sigma(\xi) = (1 - \theta)\xi + \theta.$$

As is well known (see Lambert [1]) this  $\theta$ -method is A-stable iff

$$(2) \quad \theta \leq \frac{1}{2}.$$

If  $\theta < \frac{1}{2}$  the method is of order 1.  $\theta = \frac{1}{2}$  gives the trapezoidal rule which is of order 2. Dahlquist [2] proved that A-stable methods cannot be of greater order than 2. Therefore, concerning k-step A-stable methods with  $k > 1$  we restrict ourselves to two-step methods of order 2 with  $\varrho, \sigma_2$  having no common root. These methods, normalized through  $\sum_{j=0}^2 \beta_j = 1$ , are given by

$$(3) \quad \left\{ \begin{array}{l} \varrho(\xi) = \alpha_2 \xi^2 + \alpha_1 \xi + \alpha_0, \quad \alpha_1 = 1 - 2\alpha_2, \quad \alpha_0 = -1 + \alpha_2 \\ \sigma(\xi) = \beta_2 \xi^2 + \beta_1 \xi + \beta_0, \quad \beta_1 = \frac{1}{2} + \alpha_2 - 2\beta_2, \quad \beta_0 = \frac{1}{2} - \alpha_2 + \beta_2 \\ \alpha_2 \geq \frac{1}{2}, \quad \beta_2 > \frac{1}{2}\alpha_2 \end{array} \right.$$

(see Dahlquist [3]).

Remark 1. Among A-stable methods those which are strongly stable at infinity (i.e. such that the roots of  $\sigma(\xi)$  lie in the interior of the unit disc) are preferable when solving stiff equations. In Zlamal [4] there is introduced a sub-class of (3) given by

$$(4) \quad \begin{aligned} g(\xi) &= (\frac{1}{2} + \vartheta)\xi^2 - 2\vartheta\xi - \frac{1}{2} + \vartheta, \\ &\qquad \qquad \qquad 0 < \vartheta \leq 1 \end{aligned}$$

$$\sigma(\xi) = \frac{1}{4}(1+\vartheta)^2\xi^2 + \frac{1}{2}(1-\vartheta^2)\xi + \frac{1}{4}(1-\vartheta)^2$$

and having an optimal stability at infinity. For methods (4) absolute values of the error constant and of the roots of  $\sigma$  are

$$(5) \quad |C| = \frac{1}{12} + \frac{1}{4}\vartheta^2, \quad |\xi_{1,2}| = \frac{1-\vartheta^2}{(1+\vartheta)^2}$$

We cannot minimize both quantities simultaneously. A reasonable compromise is to take  $\vartheta = \frac{1}{3}$ ; then  $|C| = \frac{1}{9}$ ,  $|\xi_{1,2}| = \frac{1}{2}$ .

Let us come back to A-stable  $\Theta$ -methods, i.e. to methods (1) with  $\Theta \leq \frac{1}{2}$ . Let  $V$  be a vector space and  $b(u, v)$  be a bilinear symmetric form on  $V \times V$ . We assume that  $b(u, v)$  is nonnegative, i.e.

$$(6) \quad 0 \leq b(u, u) \leq \|u\|^2 \quad \forall u \in V.$$

Consider the sequence

$$(7) \quad S^n = b(u^{n+1} - u^n, (1-\varepsilon)u^{n+1} + \Theta u^n)$$

$v_i, i = 1, 2, \dots, n$   $\in V$ ,  $n=1, 2, \dots$ . Then

$$\begin{aligned} \sum_{m=0}^{m-1} S^m &= \sum_{m=0}^{m-1} \left\{ \frac{1}{2} [ |u^{m+1}|^2 - |u^m|^2 ] + (\frac{1}{2} - \theta) |u^{m+1} - u^m|^2 \right\} \geq \\ &\geq \frac{1}{2} \sum_{m=0}^{m-1} [ |u^{m+1}|^2 - |u^m|^2 ]. \end{aligned}$$

Hence

$$(8) \quad \sum_{m=0}^{m-1} S^m \geq \frac{1}{2} ( |u^m|^2 - |u^0|^2 ).$$

If we apply the  $\Theta$ -method to the solution of

$$(9) \quad \frac{dx}{dt} = -\lambda x, \quad x(0) = x_0, \quad \lambda \geq 0,$$

we get for the discrete solution  $\{x_m\}_{m=0}^{\infty}$  from (8) ( $V=R^1$ ,  $b(x,y)=xy$ )

$$x_m^2 \leq x_0^2, \quad m = 1, 2, \dots .$$

The same property has the exact solution:  $x^2(t) \leq x^2(0)$  for  $t > 0$ . The energy inequality (3) (with  $V=L^2(\Omega)$ ,  $b(u,v)=(u,v)_{L^2(\Omega)}$ ) was used by many authors to derive bounds for the error of approximate solution to parabolic equations.

The energy inequality is preserved by schemes (3) as well. It was proved in Zlámal [5] in a little different form. We set

$$(10) \quad S^m = b\left(\sum_{j=0}^k \alpha_j u^{m+j}, \sum_{j=0}^k \beta_j u^{m+j}\right)$$

and denote

$$(11) \quad \mathcal{D} = \beta_2 - \frac{1}{2} \alpha_2 > 0.$$

An easy computation gives

$$\begin{aligned} S^m &= \frac{1}{2}(\alpha_2^2 + \delta)|u^{m+2}|^2 - (\alpha_2 - \frac{1}{2})|u^{m+1}|^2 - \frac{1}{2}[(\alpha_2 - 1)^2 + \delta]|u^m|^2 - \\ &\quad - [\alpha_2(\alpha_2 - 1) + \delta][b(u^{m+2}, u^{m+1}) - b(u^{m+1}, u^m)] + \\ &\quad + \delta(\alpha_2 - \frac{1}{2})|u^{m+2} - 2u^{m+1} + u^m|^2. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{n=0}^{m-1} S^n &\geq \frac{1}{2}(\alpha_2^2 + \delta)|u^m|^2 + \frac{1}{2}[(\alpha_2 - 1)^2 + \delta]|u^{m-1}|^2 - \\ &\quad - [\alpha_2(\alpha_2 - 1) + \delta]b(u^m, u^{m-1}) - R(u^0, u^1), \end{aligned}$$

$$(12) \quad |R(u^0, u^1)| \leq C(\alpha_2, \delta)(|u^0|^2 + |u^1|^2).$$

As

$$\begin{aligned} &(\alpha_2^2 + \delta)|u^m|^2 + [(\alpha_2 - 1)^2 + \delta]|u^{m-1}|^2 - \\ &- 2[\alpha_2(\alpha_2 - 1) + \delta]b(u^m, u^{m-1}) = \frac{\delta}{(\alpha_2 - 1)^2 + \delta}|u^m|^2 + \\ &+ \frac{1}{(\alpha_2 - 1)^2 + \delta}|[\alpha_2(\alpha_2 - 1) + \delta]u^m - [(\alpha_2 - 1)^2 + \delta]u^{m-1}|^2, \end{aligned}$$

it follows

$$(13) \quad \sum_{n=0}^{m-1} S^n \geq c(\alpha_2, \delta)|u^m|^2 - C(\alpha_2, \delta)(|u^0|^2 + |u^1|^2),$$

$$c(\alpha_2, \delta) = \frac{1}{2} \frac{\delta}{(\alpha_2 - 1)^2 + \delta} > 0.$$

We write the inequalities (8) and (13) in a joint form.

Let

$$(14) \quad S^m = b \left( \sum_{j=0}^k \alpha_j u^{m+j}, \sum_{j=0}^k \beta_j u^{m+j} \right), \quad k=1,2.$$

where

$$(15) \quad \begin{aligned} \alpha_1 &= 1, \quad \alpha_0 = -1, \quad \beta_1 = 1-\Theta, \quad \beta_0 = \Theta, \quad \Theta \leq \frac{1}{2} \text{ if } k=1, \\ \alpha_1 &= 1-2\alpha_2, \quad \alpha_0 = -1+\alpha_2, \quad \beta_1 = \frac{1}{2} + \alpha_2 - 2\beta_2, \quad \beta_0 = \frac{1}{2} - \alpha_2 + \beta_2, \\ \alpha_2 &\geq \frac{1}{2}, \quad \beta_2 > \frac{1}{2}\alpha_2 \quad \text{if } k=2. \end{aligned}$$

Then it holds

$$(16) \quad |u^m|^2 \leq C_1 \sum_{j=0}^{k-1} |u^j|^2 + C_2 \sum_{m=0}^{m-k} S^m, \quad m \geq k \quad (k=1,2);$$

here  $C_1, C_2$  are positive constants depending on the coefficients  $\alpha_j, \beta_j$  only.

From (16) it follows that the approximate solution of (9) satisfies the inequality

$$x_m^2 \leq C_1 \sum_{j=0}^{k-1} x_j^2, \quad m \geq k.$$

Another application of (16) concerns the problem

$$(17) \quad \frac{d^2x}{dt^2} + d \frac{dx}{dt} + \omega^2 x = 0, \quad x(0) = y_0, \quad \frac{dx(0)}{dt} = \alpha_0, \quad d, \omega = \text{const}, d \geq 0.$$

We write (17) as a system

$$(18) \quad \frac{dy}{dt} = \alpha, \quad \frac{dx}{dt} = -\omega^2 y - d\alpha,$$

we apply the method (15) and multiply the first equation by

$\omega^2 \sum_{j=0}^k \beta_j \gamma_{m+j}$  and the second by  $\sum_{j=0}^k \beta_j \alpha_{m+j}$ . Using (16) we obtain

$$(19) \quad \omega^2 \gamma_m^2 + \alpha_m^2 \leq C_1 \sum_{j=0}^{k-1} (\omega^2 \gamma_j^2 + \alpha_j^2), \quad m \geq k.$$

Remark 2. It is easy to see that if a linear multistep scheme with an arbitrary number of steps and of arbitrary order of accuracy has the property (16) then the method is A-stable (in fact, one proves that the method is A-stable in the sense of definition by Crouzeix-Raviart [6], p.40; however this definition is equivalent with the classical Dahlquist definition - see [6], p.41).

### III. Nonlinear Heat Equation

1. Let  $\Omega \subset \mathbb{R}^N$  be a bounded domain with a boundary  $\partial\Omega$ ,  $x = (x_1, \dots, x_N)$  and  $\{k_{ij}(x, t, u)\}_{i,j=1}^N$  be a uniformly positive definite matrix. Further, the coefficients  $k_{ij}(x, t, u)$  are supposed to be uniformly Lipschitz continuous functions of  $t \in [0, T]$  and of  $u \in (-\infty, \infty)$  and the right-hand side  $f(x, t, u)$  a uniformly Lipschitz continuous function of  $u \in (-\infty, \infty)$ . We consider the problem

$$\frac{\partial u}{\partial t} = \sum_{i,j=1}^N \frac{\partial}{\partial x_i} [k_{ij}(x, t, u) \frac{\partial u}{\partial x_j}] + f(x, t, u) \text{ in } \Omega \times (0, T)$$

$$u(x, t) = 0 \text{ on } \partial\Omega \times (0, T), \quad 0 < T < \infty,$$

$$u(x, 0) = u^0(x) \text{ in } \Omega.$$

More general equations and boundary conditions can be treated in the same way we will now describe.

If the exact solution is smooth enough then it holds

$$(21) \quad (u', v)_0 + a(t, u; u, v) = (f(x, t, u), v)_0 \text{ in } [0, T] \quad \forall v \in H_0^1(\Omega).$$

Here

$$u' = \frac{\partial u}{\partial t}, \quad (u, v)_0 = \int_{\Omega} uv dx,$$

$$a(t, w; u, v) = \int_{\Omega} \sum_{i,j=1}^N b_{ij}(x, t, w) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx,$$

$H^m(\Omega)$  is the Sobolev space  $\{u \in L^2(\Omega); D^\alpha u \in L^2(\Omega) \quad \forall |\alpha| \leq m\}$

with the usual scalar product  $(u, v)_m = \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}$  and

the norm  $\|u\|_m = (u, u)_m^{\frac{1}{2}}$  and  $H_0^1(\Omega) = \{u \mid u \in H^1(\Omega); u|_{\partial\Omega} = 0\}$ .

Let us consider a family of finite element spaces such that  $V^h \subset H_0^1(\Omega)$ . The Galerkin method yields a semidiscrete solution  $U(x, t)$  which for each  $t \in [0, T]$  is a function from  $V^h$ .  $U(x, t)$  is uniquely determined by a discrete analog of (21):

$$(22) \quad (U', v)_0 + a(t, U; U, v) = (f(x, t, U), v)_0 \quad \forall v \in V^h,$$

$$U(x, 0) = U^0(x);$$

$U^0(x)$  is a suitable approximation of  $u^0(x)$  from  $V^h$ . (22) represents a system of ordinary differential equations. Applying the method (15) we obtain

$$(23) \quad \begin{aligned} & \left( \sum_{j=0}^k \alpha_j U^{m+j}, v \right)_0 + \Delta t \sum_{j=0}^k \beta_j a(t_{m+j}, U^{m+j}; U^{m+j}, v) = \\ & = \Delta t \sum_{j=0}^k \beta_j (f(x, t_{m+j}, U^{m+j}), v)_0 \quad \forall v \in V^h, \quad 0 \leq m \leq [\frac{T}{\Delta t}] - k. \end{aligned}$$

The scheme (23) being nonlinear has little practical value. We linearize it as follows:

$$(24) \quad \left( \sum_{j=0}^k \alpha_j U^{m+j}, v \right)_0 + \Delta t \sum_{j=0}^k \beta_j a(t_{\bar{n}}, U^{\bar{n}}; U^{m+j}, v) = \Delta t (f(x, t_{\bar{n}}, U^{\bar{n}}), v)_0 \\ \forall v \in V^h, \quad 0 \leq m \leq [\frac{T}{\Delta t}] - k;$$

for  $k=1$

$$(25) \quad t^{\bar{n}} = \begin{cases} t_m, & \theta < \frac{1}{2} \\ t_m + \frac{1}{2} \Delta t, & \theta = \frac{1}{2} \end{cases}, \quad U^{\bar{n}} = \begin{cases} U^m, & \theta < \frac{1}{2} \\ \frac{3}{2} U^m - \frac{1}{2} U^{m-1}, & \theta = \frac{1}{2} \end{cases}$$

(see Douglas and Dupont [7]),

for  $k=2$

$$(26) \quad t_{\bar{n}} = t_m + (\frac{1}{2} + \alpha_2) \Delta t, \quad U^{\bar{n}} = (\frac{1}{2} + \alpha_2) U^{m+1} + (\frac{1}{2} - \alpha_2) U^m$$

(see Zlámal [5]). The order of accuracy  $q$  of the method (24) is equal 1 if  $k=1$  and  $\theta < \frac{1}{2}$  and 2 if  $k=1$ ,  $\theta = \frac{1}{2}$  or  $k=2$ .

Notice that whereas (23) with  $k=1$  and  $\theta = \frac{1}{2}$  is a one-step scheme the corresponding scheme (24) is a two-step scheme.

Remark 3. Even when (24) represents a linear algebraic system at every time step it is not the final scheme in practical computations. In general, we have to consider finite element spaces  $V^h$  which are subspaces of  $H_0^1(\Omega_h)$ ,  $\Omega_h \neq \Omega$  (best known example: curved isoparametric elements). In addition, we have to compute mass and stiffness matrices numeri-

callly. Let us denote by  $(u, v)_h$  and  $a_h(t, w; u, v)$  the approximate values of

$$\int_{\Omega_h} uv dx \quad \text{and} \quad \sum_{k=1}^N \sum_{i,j=1}^N b_{ij}(x, t, w) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx,$$

respectively, computed by a suitable quadrature rule. The final scheme is

$$(27) \quad \begin{aligned} & \left( \sum_{j=0}^k \alpha_j U^{m+\delta}, v \right)_h + \Delta t \sum_{j=0}^k \beta_j a_h(t_{\bar{n}}, U^{\bar{n}}; U^{m+\delta}, v) = \\ & = \Delta t (f(x, t_{\bar{n}}, U^{\bar{n}}), v)_h \quad \forall v \in V^h, \quad 0 \leq m \leq [\frac{T}{\Delta t}] - k. \end{aligned}$$

When possible we use for the computation of  $(u, v)_{L^2(\Omega_h)}$  a quadrature formula such that the (mass) matrix corresponding to  $(., .)_h$  be diagonal. The engineers speak about lumping (see Zienkiewicz [8], p.525).

2. We outline the way how to derive error bounds. We assume that the family  $\{V^h\}$ ,  $0 < h < h^*$ , has the following approximation property shared by finite element subspaces: to any  $u \in H^{p+1}(\Omega) \cap H_0^1(\Omega)$  there exists  $u^h \in V^h$  such that

$$(28) \quad \|u - u^h\|_0 + h \|u - u^h\|_1 \leq C h^{p+1} \|u\|_{p+1}.$$

(in the sequel,  $C$  denotes a constant not necessarily the same at any two places which may depend on  $u$ ). We decompose the exact

solution in  $u = \xi + \eta$  where  $\xi \in V^h$  is the Ritz approximation defined by

$$(29) \quad a(t, u; u, v) = a(t, u; \xi, v) \quad \forall v \in V^h.$$

Under some assumptions one can prove that

$$(30) \quad \|\eta\|_0 + \|\eta'\|_0 \leq Ch^{k+1} [\|u\|_{p+1} + \|u'\|_{p+1}] \quad \forall t \in (0, T]$$

(see Wheeler [9] and Dupont, Fairweather, Johnson [10]). Hence, it is sufficient to estimate  $e^m = \xi^m - U^m$ ,  $\xi^m = \xi(t_m)$ . One derives (see Zlámal [5]) that

$$(31) \quad \left( \sum_{j=0}^k \alpha_j e^{m+j}, v \right)_0 + \Delta t a(t_m, U^{\bar{m}}; \sum_{j=0}^k \beta_j e^{m+j}, v) = \Delta t (\psi^m, v), \quad \forall v \in V^h,$$

$$(32) \quad \|\psi^m\|_1 \leq C(h^{k+1} + \Delta t^q + \|\varepsilon^{\bar{m}}\|);$$

(again,  $q=1, 2$  is the order of accuracy). We choose  $v = \sum_{j=0}^k \beta_j e^{m+j}$  in (31) and use the uniform  $H^1(\Omega)$ -ellipticity of the form  $a(t, w; u, v)$  following from the uniformly positive definiteness of the matrix  $\{k_{ij}(x, t, w)\}$ , i.e.  $a(t, w; v, v) \geq \beta \|v\|_1^2 \quad \forall v \in H_0^1(\Omega)$ ,

$\beta > 0$ . Estimating the right-hand side of (31) by  $|(\psi^m, v)| \leq \frac{1}{2} \delta \|\psi^m\|_1^2 + \frac{1}{2\delta} \|v\|_1^2$  with a suitable  $\delta$  and taking into account (32) we get  $(b(u, v) = (u, v))$

$$S^m + \beta \Delta t \left\| \sum_{j=0}^k \beta_j e^{m+j} \right\|_1^2 \leq \frac{1}{2} \beta \Delta t \left\| \sum_{j=0}^k \beta_j e^{m+j} \right\|_1^2 + C \Delta t ([h^{k+1} + \Delta t^q]^2 + \|\varepsilon^{\bar{m}}\|_0^2).$$

(16) gives

$$\|e^m\|_0^2 \leq C_1 \sum_{j=0}^{k-1} \|e^j\|_0^2 + C [h^{k+1} + \Delta t^q]^2 + C \Delta t \sum_{j=0}^{m-1} \|e^j\|_0^2.$$

The discrete Gronwall inequality (see Lees [11]) implies

$$\|e^m\|_0^2 \leq C \left\{ \sum_{j=0}^{k-1} \|e_j\|_0^2 + [\hbar^{p+1} + \Delta t^q]^2 \right\}$$

from which using (30) one gets easily the final result

$$(33) \quad \|u^m - U^m\|_0 \leq C \left\{ \sum_{j=0}^{k-1} \|u_j - U_j\|_0 + \hbar^{p+1} + \Delta t^q \right\}.$$

#### IV. Time Dependent Navier-Stokes Equations

1. Whereas in the preceding section we did not precise the variational formulation of the problem we want to do it here. To this end we introduce some spaces of functions valued in a Banach space, we define the weak or generalized derivative of such functions and consider a certain space suitable to the solution of time dependent problems.

Let  $X$  be a Banach space normed by  $\|\cdot\|_X$  and let

$$0 < T < \infty.$$

For  $p \geq 1$  we denote by  $L^p(0, T; X)$  the space of strongly measurable functions  $f: (0, T) \rightarrow X$  (see, e.g. Kufner+John+Fucik [12], p.107) such that

$$\|f\|_{L^p(0, T; X)} = \left[ \int_0^T \|f(t)\|_X^p dt \right]^{\frac{1}{p}} < \infty \quad \text{if } 1 \leq p < \infty,$$

$$\|f\|_{L^\infty(0, T; X)} = \operatorname{ess\,sup}_{t \in (0, T)} \|f(t)\|_X < \infty \quad \text{if } p = \infty.$$

By  $C([0,T];X)$  we denote the space of continuous functions  $f: [0,T] \rightarrow X$  normed by

$$\|f\|_{C([0,T];X)} = \max_{t \in [0,T]} \|f(t)\|_X.$$

To define the weak or generalized derivative of a function valued in a Banach space we introduce the following

**Lemma 1.** Let  $X$  be a given Banach space,  $X'$  its dual and let  $u$  and  $g$  be two functions belonging to  $L^1(0,T;X)$ . Then the following three conditions are equivalent:

i)  $u$  is a.e. equal to a primitive function of  $g$ ,

$$u(t) = \xi + \int_0^t g(s) ds, \quad \xi \in X, \quad \text{a.e. in } (0,T),$$

(all integrals with respect to the time are Bochner integrals; see, e.g., Kufner+John+Fucík [12], sect.2.19),

$$\text{ii) } \int_0^T u(t) \varphi'(t) dt = - \int_0^T g(t) \varphi(t) dt \quad \forall \varphi \in \mathcal{D}((0,T)),$$

iii) for all  $\eta \in X'$

$$\frac{d}{dt} \langle \eta, u \rangle = \langle \eta, g \rangle \quad \text{in } \mathcal{D}'((0,T))$$

where  $\langle \dots \rangle$  is the scalar product in the duality between  $X'$  and  $X$ . In addition, in each of these cases  $u$  is a.e. equal to a function of  $C([0,T];X)$ .

The proof of Lemma 1 can be found in Temman [13], p.250. The function  $g$  of this lemma - the weak or generalized derivative of  $u$ -is denoted by  $u'$  or  $\frac{du}{dt}$ .

If  $u \in L^1(0, T; X)$  is a solution of an evolution equation which should satisfy the initial condition  $u(0) = u_0$  and if we find out that from the equation it follows  $u' \in L^1(0, T; X)$  then according to lemma 1 it holds  $u \in C([0, T]; X)$  and the initial condition makes sense if  $u_0 \in X$  and if we take it as  $\lim_{t \rightarrow 0^+} \|u(t) - u_0\|_X = 0$ .

Often, we have a different situation. Let us consider the simple problem  $\frac{\partial u}{\partial t} - \Delta u = 0$  in  $\Omega$ ,  $u|_{\partial\Omega} = 0$ ,  $u(x, 0) = u_0(x)$  in  $\Omega$  as an operator equation. Taking  $-\Delta u$  in the distributional sense we have  $\langle -\Delta u, \varphi \rangle = \sum_{i=1}^N \langle \frac{\partial u}{\partial x_i}, \frac{\partial \varphi}{\partial x_i} \rangle$ . If  $u \in H_0^1(\Omega)$  then the right-hand side is bounded by  $\|u\|_{H_0^1(\Omega)} \|\varphi\|_{H_0^1(\Omega)}$ . Hence, for  $u \in L^2(0, T; H_0^1(\Omega))$ ,  $-\Delta u$  maps  $L^2(0, T; H_0^1(\Omega))$  into  $L^2(0, T; H^{-1}(\Omega))$  ( $H^{-1}(\Omega)$  is the dual of  $H_0^1(\Omega)$ ). From the equation it follows that  $\frac{\partial u}{\partial t} \in L^2(0, T; H^{-1}(\Omega))$ . More generally, let us assume that there are given a Hilbert space  $H$  with a scalar product  $(\cdot, \cdot)$  and norm  $\|\cdot\|$  and a reflexive Banach space  $V$  with a norm  $\|\cdot\|$  which is dense and continuously imbedded in  $H$  (in case of the heat equation  $H = L^2(\Omega)$ ,  $V = H_0^1(\Omega)$ ). We identify  $H$  with its dual space by means of its scalar product. Then  $H$  can be identified with a subspace of  $V'$  and we have inclusions

$$(34) \quad V \subset H \subset V'$$

where each space is dense in the following one and the injections are continuous. Furthermore, the scalar product  $\langle \cdot, \cdot \rangle$  between  $V'$  and  $V$  is an extension of  $(\cdot, \cdot)$ . Now, let us consider an operator equation  $u' + A(u) = f$  with the initial condition  $u(0) = u_0$ .  $A(u)$  is supposed to be a nonlinear operator from  $L^p(0, T; V)$  into  $L^{p'}(0, T; V')$ ,  $\frac{1}{p} + \frac{1}{p'} = 1$ , and  $f \in L^{p'}(0, T; V')$ . Looking for  $u \in L^p(0, T; V)$  we see from the equation that

$u' \in L^p(0, T; V')$ . The following lemma guarantees that the initial condition  $U(0) = u_0$  makes sense if we assume  $u_0 \in H$  and if we take it as  $\lim_{t \rightarrow 0^+} [u(t) - u_0] = 0$ .

**Lemma 2.** Let  $H$  be a Hilbert space and  $V$  a reflexive Banach space which is dense and continuously imbedded in  $H$ . Let  $W$  be the Banach space  $W = \{v/v \in L^p(0, T; V); v' \in L^{p'}(0, T; V')\}$ ,  $1 < p < \infty$ ,  $\frac{1}{p} + \frac{1}{p'} = 1$ , normed by  $\|v\|_W = \|v\|_{L^p(0, T; V)} + \|v'\|_{L^{p'}(0, T; V')}$ . Then  $W \subset C([0, T]; H)$  and the imbedding is continuous. Furthermore, for any  $u, v \in W$  it holds the formula of integration by parts

$$(35) \quad \int_0^t \{ \langle u', v \rangle + \langle v', u \rangle \} d\tau = (u(t), v(t)) - (u(0), v(0)), \quad 0 < t \leq T.$$

The lemma is true even in a somewhat more general form and the proof can be found in Gajewski+Gröger+Zacharias [13], p.147.

2. The classical formulation of the initial boundary value problem to the Navier-Stokes equations is the following: Find a vector function  $\underline{u} = (u_1(x, t), \dots, u_N(x, t))^T$  ( $T$  written as a superscript denotes transposition of a vector or of a matrix) and a scalar function  $p(x, t)$  such that

$$(36) \quad \frac{\partial}{\partial t} \underline{u} - \nu \Delta \underline{u} + \sum_{j=1}^N u_j \frac{\partial}{\partial x_j} \underline{u} + \text{grad } p = \underline{f}(x, t), \quad \left. \right\} \text{in } \Omega \times (0, T)$$

$$(37) \quad \text{div } \underline{u} = \underline{g},$$

$$(38) \quad \underline{u} = \underline{0} \quad \text{on } \partial\Omega \times (0, T),$$

$$(39) \quad \underline{u}(x, 0) = \underline{u}_0(x) \quad \text{in } \Omega.$$

In these equations  $\Omega \subset \mathbb{R}^N$  is a bounded domain with the boundary  $\partial\Omega$ ,  $x = (x_1, \dots, x_N)$ , the vector function  $\underline{u}$  is the velocity of the  $N$ -dimensional motion of a viscous incompressible fluid,  $p$  is the kinematic pressure,  $\nu > 0$  assumed to be constant is the kinematic viscosity,  $f$  represents a density of body forces per unit mass and  $\underline{u}_0$  is the initial velocity. We restrict ourselves to two and three dimensions:

$$N = 2, 3.$$

Evidently, we have to consider now the spaces  $(L^2(\Omega))^N$  and  $(H^1(\Omega))^N$  with the usual scalar product and norm which we denote as in case  $N=1$  (see sect. III):

$$(\underline{u}, \underline{v})_m = \sum_{i=1}^N (u_i, v_i)_m, \quad \|\underline{u}\|_m = \left\{ \sum_{i=1}^N \|u_i\|_m^2 \right\}^{\frac{1}{2}}, \quad m=0,1.$$

Let  $V$  be the space (without topology)

$$(40) \quad V = \{ \underline{u} \in (\mathcal{D}(\Omega))^N ; \operatorname{div} \underline{u} = 0 \}.$$

The closures  $H$  and  $V$  of  $V$  in  $(L^2(\Omega))^N$  and  $(H^1(\Omega))^N$ , respectively, are basic spaces in the study of the Navier-Stokes equations. It is proved in Teman [13], p.15 and 18, that if  $\partial\Omega$  is a Lipschitz boundary then

AD-A110 966

MARYLAND UNIV COLLEGE PARK DEPT OF MATHEMATICS  
LECTURES ON THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUA--ETC(U)  
DEC 81 I BABUSKA, T - LIU, J OSBORN AFOSR-80-0251

F/G 12/1

UNCLASSIFIED

AFOSR-TR-82-0047

NL

7 x 7  
4. Other

END  
DATA  
FILED  
3. 82  
DTIC

$$(41) \quad \left\{ \begin{array}{l} H = \{ \underline{u} \in (L^2(\Omega))^N; \operatorname{div} \underline{u} = 0; \gamma_N \underline{u} = 0 \}, \\ V = \{ \underline{v} \in (H_0^1(\Omega))^N; \operatorname{div} \underline{v} = 0 \}. \end{array} \right.$$

Here  $\gamma_N \underline{u} = \underline{u} \cdot \underline{v}|_{\partial\Omega} = \sum_{i=1}^N u_i v_i|_{\partial\Omega}$  and  $\underline{v}$  is the unit exterior normal to  $\partial\Omega$ . As  $\|\underline{u}\|_0 \leq \|\underline{u}\|_1$ ,  $V$  is dense and continuously imbedded in  $H$ , hence  $H$  and  $V$  are examples of abstract spaces introduced in paragraph 1, we have inclusions (34) and the scalar product  $\langle \dots \rangle$  between  $V'$  and  $V$  is an extension of  $\langle \dots \rangle_0$ .

To give a variational formulation of the problem (36)-(39) let us first consider sufficiently smooth functions  $\underline{u}, p$  say  $\underline{u} \in (C^2(\bar{\Omega} \times [0,T]))^N$ ,  $p \in C^1(\bar{\Omega} \times [0,T])$ , satisfying the equations (36), (37) and (38). Certainly,  $\underline{u}$  belongs to  $L^2(0,T;V)$ . Further, multiplying (36) by a function  $\underline{v} \in V$  and integrating we get

$$\frac{d}{dt} (\underline{u}, \underline{v})_0 - v(\Delta \underline{u}, \underline{v})_0 + a_1(\underline{u}; \underline{u}, \underline{v}) + (\underline{\operatorname{grad}} p, \underline{v})_0 = (\underline{f}, \underline{v})_0$$

where

$$a_1(\underline{w}; \underline{u}, \underline{v}) = \sum_{i,j=1}^N \int_{\Omega} w_j \frac{\partial u_i}{\partial x_j} v_i dx.$$

Using Green's theorem we obtain  $-v(\Delta \underline{u}, \underline{v})_0 = a_0(\underline{u}, \underline{v})$  with

$$a_0(\underline{u}, \underline{v}) = v \int_{\Omega} \sum_{i,j=1}^N \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} dx$$

and

$$(\underline{\operatorname{grad}} p, \underline{v})_0 = 0$$

(due to  $\underline{v}/\partial \Omega = 0$ ,  $\gamma_v \underline{v} = 0$ ,  $\operatorname{div} \underline{v} = 0$ ). Denoting

$$(42) \quad a(\underline{w}; \underline{u}, \underline{v}) = a_0(\underline{u}, \underline{v}) + a_1(\underline{w}; \underline{u}, \underline{v})$$

we see that

$$\frac{d}{dt} (\underline{u}, \underline{v})_0 + a(\underline{u}; \underline{u}, \underline{v}) = (\underline{f}, \underline{v})_0 \quad \forall \underline{v} \in V.$$

By continuity this equation is true for each  $\underline{v} \in V$ .

The preceding lines suggest the following variational formulation of the problem (36)-(39): For a given right-hand side  $\underline{f} \in L^2(0, T; (H^{-1}(\Omega))^N)$  and a given initial value  $\underline{u}_0 \in H$  find  $\underline{u} \in L^2(0, T; V)$  such that

$$(43) \quad \frac{d}{dt} (\underline{u}, \underline{v})_0 + a(\underline{u}; \underline{u}, \underline{v}) = \langle \underline{f}, \underline{v} \rangle \quad \text{in } D'((0, T)) \quad \forall \underline{v} \in V,$$

$$(44) \quad \underline{u}(0) = \underline{u}_0.$$

It is proved in Girault-Raviart [15] (chap. V, theorem 1.4 · 1.2 and 1.5) that there is a function  $\underline{u}$  satisfying (43) which lies even in a smaller space:  $\underline{u} \in L^2(0, T; V) \cap L^\infty(0, T; H)$ . In addition,  $\dot{\underline{u}} \in L^2(0, T; V')$  if  $N=2$  and  $\dot{\underline{u}} \in L^{\frac{4}{3}}(0, T; V')$  if  $N=3$ . Therefore, the initial condition (44) makes sense and is satisfied in the following form:  $\lim_{t \rightarrow 0^+} \|\underline{u}(t) - \underline{u}_0\|_0 = 0$  and  $\lim_{t \rightarrow 0^+} \|\dot{\underline{u}}(t) - \dot{\underline{u}}_0\|_{V'} = 0$ , respectively. Finally, if  $N=2$  such a solution is unique.

3. We define a semidiscrete solution of the problem (43), (44) applying the scheme (24)-(26) and derive error bounds by

means of the inequality (16). These results belong to Girault + Raviart [15] (chap.V, §3).

We remind the reader that the method (24)-(26) is of order  $q=1$  iff  $k=1$  and  $\theta < \frac{1}{2}$  and of order  $q=2$  iff  $k=1$  and  $\theta = \frac{1}{2}$  or  $k=2$ . The approximate value of  $\underline{u}(t_n) = \underline{u}^n(t_n = n\Delta t)$  is denoted by  $\underline{u}^n$  and recurrently defined as follows:

$$(45) \quad \begin{aligned} \underline{U}^n \in V, \quad 0 \leq n \leq [\frac{T}{\Delta t}] = M, \\ (\sum_{j=0}^k \alpha_j \underline{U}^{n+j}, \underline{v})_0 + \Delta t \sum_{j=0}^k \beta_j a(\underline{U}^{\bar{n}}; \underline{U}^{n+j}, \underline{v}) = \Delta t \langle \underline{f}, \underline{v} \rangle_0 \\ \forall \underline{v} \in V, \quad 0 \leq n \leq M-k, \end{aligned}$$

$$(46) \quad \begin{cases} \underline{U}^0 = \underline{u}_0 & \text{if } q=1, \\ \underline{U}^0 = \underline{u}_0, \quad \underline{U}^1 = \underline{u}_1 & \text{if } q=2. \end{cases}$$

Here  $t_{\bar{n}}$  and  $\underline{U}^{\bar{n}}$  are defined by (25) and (26),  $\underline{f}^{\bar{n}} = \underline{f}(x, t_{\bar{n}})$ ,  $\underline{u}_0 = \underline{u}(0)$  is now supposed to lie in  $V$  and  $\underline{u}_1 \in V$  is given. Of course,  $\underline{u}_1$  should be an enough accurate approximation of  $\underline{u}(t_1)$ . We can take for  $\underline{u}_1$  the value  $\underline{U}^1$  computed by the  $\theta$ -method with  $\theta < \frac{1}{2}$ .

Given the starting values the equation (45) defines a unique set  $\{\underline{U}^n\}_{n=0}^M$ . To see it we remark that the function  $\underline{U}^{n+k}$  is the solution in  $V$  of the linear boundary value problem

$$(47) \quad \alpha_k(\underline{U}^{n+k}, \underline{v})_0 + \beta_k \Delta t a(\underline{U}^{\bar{n}}; \underline{U}^{n+k}, \underline{v}) = \langle \underline{f}^{\bar{n}}, \underline{v} \rangle \quad \forall \underline{v} \in V$$

where  $\underline{f}^{\bar{n}}$  is a known element of  $V'$ . The trilinear form  $a(\underline{w}; \underline{u}, \underline{v})$  is continuous on  $(V)^N$ ,  $N \leq 4$  (see lemma 2.1, p.114 in [15]). Also, it is (uniformly)  $V$ -elliptic as a bilinear form in  $\underline{u}, \underline{v}$ :

$$(48) \quad a(\underline{w}; \underline{v}, \underline{v}) \geq \alpha \|\underline{v}\|_1^2, \quad \forall \underline{v}, \underline{w} \in V, \quad \alpha = \text{const} > 0.$$

We have namely

$$\begin{aligned} a_1(\underline{w}; \underline{v}, \underline{v}) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \int_{\Omega} w_j \frac{\partial}{\partial x_j} (v_i^2) dx = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \int_{\Omega} v_i^2 \frac{\partial w_j}{\partial x_j} dx = \\ &= -\frac{1}{2} \sum_{i=1}^N \int_{\Omega} v_i^2 \operatorname{div} \underline{w} dx = 0. \end{aligned}$$

Therefore

$$a(\underline{w}; \underline{v}, \underline{v}) = a_0(\underline{v}, \underline{v}) = V \sum_{i=1}^N \int_{\Omega} \sum_{j=1}^N \left( \frac{\partial v_i}{\partial x_j} \right)^2 dx.$$

Applying Friedrichs inequality we get (48) (of course,  $\alpha$  is a multiple of  $V$ ). As  $\alpha_2, \beta_2$  are positive the form  $a_2(\underline{u}, \underline{v})_0 + \beta_2 \Delta t a(\underline{U}^n; \underline{u}, \underline{v})$  is  $V$ -elliptic. This proves existence and uniqueness of  $\{\underline{U}^n\}_{n=1}^M$ .

Let us now introduce the regularity conditions for the exact solution  $\underline{u}$ . We assume

$$(49) \quad \underline{u} \in C([0, T]; (L^\infty(\Omega))^N),$$

$$(50) \quad \underline{u}' \in L^2(0, T; V), \quad \underline{u}'' \in L^2(0, T; V') \quad \text{if } q=1,$$

$$(51) \quad \underline{u}'' \in L^2(0, T; V), \quad \underline{u}''' \in L^2(0, T; V') \quad \text{if } q=2.$$

In view of lemma 1 and 2 the condition (50) implies that

$\underline{u}' \in C([0, T]; V)$  and  $\underline{u}'' \in C([0, T]; H)$ . As a consequence,

$a(\underline{u}(t); \underline{u}(t), \underline{y}) \in C([0, T])$ ,  $\frac{d}{dt}(\underline{u}(t), \underline{y})_0 = (\underline{u}'(t), \underline{y})_0 \in C([0, T])$ , hence due to (43)  $\langle \underline{f}(t), \underline{y} \rangle \in C([0, T])$  and

$$(52) \quad (\underline{u}'(t), \underline{y})_0 + a(\underline{u}(t); \underline{u}(t), \underline{y}) = \langle \underline{f}(t), \underline{y} \rangle \quad \text{in } [0, T] \quad \forall \underline{y} \in V.$$

We want to estimate

$$\underline{e}^n = \underline{u}(t_n) - \underline{U}^n \equiv \underline{u}^n - \underline{U}^n.$$

We define the truncation error by

$$(53) \quad \Delta t \langle \underline{\xi}^n, \underline{v} \rangle = \left( \sum_{j=0}^k \alpha_j \underline{u}^{n+j}, \underline{v} \right)_0 + \Delta t a(\underline{y}^{\bar{n}}; \sum_{j=0}^k \beta_j \underline{u}^{n+j}, \underline{v}) - \Delta t \langle \underline{f}^{\bar{n}}, \underline{v} \rangle \quad \forall \underline{v} \in V.$$

From (45) and (53) it easily follows

$$(54) \quad \left( \sum_{j=0}^k \alpha_j \underline{e}^{n+j}, \underline{v} \right)_0 + \Delta t a(\underline{U}^{\bar{n}}; \sum_{j=0}^k \beta_j \underline{e}^{n+j}, \underline{v}) = \Delta t \langle \underline{\xi}^n, \underline{v} \rangle - \Delta t a_1(\underline{e}^{\bar{n}}; \sum_{j=0}^k \beta_j \underline{U}^{n+j}, \underline{v}) \quad \forall \underline{v} \in V.$$

Before applying the inequality (16) we have to estimate the terms on the right-hand side of (54). Choosing suitably  $\delta$  in the inequality  $a b \leq \frac{1}{2} (\delta a^2 + \delta^{-1} b^2)$  we have

$$(55) \quad | \langle \underline{\xi}^n, \underline{v} \rangle | \leq \frac{1}{4} \alpha \| \underline{v} \|_1^2 + C \| \underline{\xi}^n \|_*^2,$$

( $\| \cdot \|_* = \| \cdot \|_V$ ) Here (and in the sequel of this section)  $C$  depend on  $\underline{u}$ . Further, for the truncation error one derives using Taylor's formula and (52) (see [15], lemma 2.6, p.178 and lemma 3.2, p.186; the case  $0 < \theta \leq \frac{1}{2}$  is not covered but can be proved in the same way) that

$$(56) \quad \Delta t \sum_{m=0}^{M-h} \| \underline{\xi}^m \|_*^2 \leq C \Delta t^{2q}, \quad q = 1, 2.$$

Concerning the form  $a_1$  we easily get using lemma 2.2 from [15], p.114 that

$$\begin{aligned} |a_1(\underline{e}^{\bar{n}}; \sum_{j=0}^k \beta_j \underline{u}^{n+j}, \underline{v})| &= |a_1(\underline{e}^{\bar{n}}; \underline{v}, \sum_{j=0}^k \beta_j \underline{u}^{n+j})| \leq \\ &\leq C \|\underline{e}^{\bar{n}}\|_0 \|\underline{v}\|_1 \end{aligned}$$

due to the regularity condition (49). Hence,

$$(57) \quad |a_1(\underline{e}^{\bar{n}}; \sum_{j=0}^k \beta_j \underline{u}^{n+j}, \underline{v})| \leq \frac{1}{4} \alpha \|\underline{v}\|_1^2 + C \|\underline{e}^{\bar{n}}\|_0^2.$$

Now, choosing  $b(u, v) = (\underline{u}, \underline{v})_0$  (see section II), putting  $\underline{v} = \sum_{j=0}^k \beta_j \underline{e}^{n+j}$  in (54) and using (48), (55), (57) we get

$$S^n + \frac{1}{2} \alpha \Delta t \left\| \sum_{j=0}^k \beta_j \underline{e}^{n+j} \right\|_1^2 \leq C \Delta t \left\{ \|\underline{e}^n\|_\infty^2 + \|\underline{e}^{\bar{n}}\|_0^2 \right\}.$$

From (16) it follows in case  $q=1$  (notice that  $\underline{e}^0=0$ )

$$(58) \quad \|\underline{e}^m\|_0^2 \leq C \Delta t \sum_{n=0}^{m-1} \|\underline{e}^n\|_\infty^2 + C \Delta t \sum_{n=0}^{m-1} \|\underline{e}^n\|_0^2 - \frac{1}{2} \alpha \Delta t \sum_{n=0}^{m-1} \left\| \sum_{j=0}^k \beta_j \underline{e}^{n+j} \right\|_1^2.$$

Thus by (56)

$$\|\underline{e}^m\|_0^2 \leq C \Delta t^2 + C \Delta t \sum_{n=0}^{m-1} \|\underline{e}^n\|_0^2$$

and by the discrete Gronwall inequality

$$(59) \quad \max_{1 \leq m \leq M} \|\underline{u}^m - \underline{U}^m\|_0 \leq C \Delta t.$$

If  $q=2$  we get in the same way

$$(60) \max_{2 \leq m \leq M} \|\underline{u}^m - \underline{U}^m\|_0 \leq C \{ \|\underline{u}(t_1) - \underline{u}_1\|_0 + \Delta t^2 \}.$$

From (58) we can also derive bounds for  $\Delta t \sum_{m=0}^{M-1} \left\| \sum_{j=0}^k \beta_j e^{(m+j)\delta} \right\|_1^a$  which are of interest in case of the Euler implicit scheme (the  $\Theta$ -method with  $\Theta = 0$ ) and of the scheme (3) with  $\alpha_2 = \frac{3}{2}$ ,  $\beta_2 = 1$  (then  $\beta_1 = \beta_0 = 0$ ). These schemes are the only two members of the backward differentiation schemes (see Lambert [1], p.242) which are A-stable. We easily derive

$$(61) \quad \left\{ \Delta t \sum_{m=1}^M \|\underline{u}^m - \underline{U}^m\|_1^2 \right\}^{\frac{1}{2}} \leq C \Delta t$$

and

$$(62) \quad \left\{ \Delta t \sum_{m=2}^M \|\underline{u}^m - \underline{U}^m\|_1^2 \right\}^{\frac{1}{2}} \leq C \{ \|\underline{u}(t_1) - \underline{u}_1\|_0^2 + \Delta t^2 \}.$$

Remark 4. It is proved in [15] (see theorem 2.2, p.179) that in case of the Euler implicit method the bounds (59) and (61) are true without assuming (49). Girault and Raviart apply a different way of estimating the form  $a_1$  for which the only regularity condition (50) is sufficient. In fact, the same trick can be used in case of the other backward differentiation scheme mentioned above and (60) and (62) are true under the regularity conditions (50) and (51) only.

## V. Nonlinear Quasistationary Magnetic Field

1. In recent years attention has been paid in electrical engineering journals to the computation of nonlinear quasistationary magnetic field. This problem occurs, e.g., in designing the magnet systems for fusion reactors and in rotating machinery. In two dimensions it can be formulated in the following model way: There is given a two-dimensional bounded domain  $\Omega$  and an open nonempty set  $R \subset \Omega$ . We are looking for a function  $u = u(x_1, x_2, t)$  (magnetic vector potential) such that

$$1) \quad \sigma \frac{\partial u}{\partial t} = \frac{\partial}{\partial x_1} (v \frac{\partial u}{\partial x_1}) + \frac{\partial}{\partial x_2} (v \frac{\partial u}{\partial x_2}) + J \quad \text{in } R,$$

$$(64) \quad u(x_1, x_2, 0) = u_0(x_1, x_2) \quad \text{in } R,$$

$$2) \quad 0 = \frac{\partial}{\partial x_1} (v \frac{\partial u}{\partial x_1}) + \frac{\partial}{\partial x_2} (v \frac{\partial u}{\partial x_2}) + J \quad \text{in } S = \Omega - \bar{R},$$

3)  $u$  satisfies a boundary condition on  $\partial\Omega$ ,

4)  $u$  satisfies the conditions

$$(66) \quad [u]_R^S = [v \frac{\partial u}{\partial n}]_R^S = 0 \quad \text{on } \Gamma = \partial R \cap \partial S.$$

Here the conductivity  $\sigma = \sigma(x_1, x_2)$  is a positive function on  $R$ , the reluctivity  $v = v(x_1, x_2, \|\operatorname{grad} u\|)$ ,  $\|\operatorname{grad} u\|^2 = (\frac{\partial u}{\partial x_1})^2 + (\frac{\partial u}{\partial x_2})^2$ , is a positive function on  $\Omega \times [0, \infty)$ ,  $J = J(x_1, x_2, t)$  is a given current density,  $u_0(x_1, x_2)$  is a given function defined on  $R$  and  $n$  is the normal oriented in a unique way.

The problem 1) - 4) can be easily formulated in a variational form. Let us, for simplicity, consider the Dirichlet boundary condition

$$(67) \quad u = 0 \quad \text{on } \partial\Omega.$$

We multiply (63) and (65) by a function  $v \in H_0^1(\Omega)$ , we integrate, we use Green's formula and (66) and we sum. The result is

$$(68) \quad (\sigma \frac{\partial u}{\partial t}, v)_{L^2(\Omega)} + a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega)$$

where

$$(69) \quad a(u, v) = \int_{\Omega} \sum_{i=1}^2 v \frac{\partial u}{\partial x_i} \frac{\partial \sigma}{\partial x_i} dx.$$

(68) is taken in Melkes-Zlámal [16] as the starting point for the construction of the approximate solution.

Here we outline main results of the paper [17]. We give two equivalent abstract formulations of the above problem. One of them is a variational formulation generalizing the special case (68). We introduce an existence and uniqueness theorem. We define a completely discretized approximate solution. The discretization in time is carried out by two members of the backward differentiation schemes mentioned at the end of sect. IV. We close this section by introducing results concerning convergence of the approximate solution and error estimates.

2. To formulate the problem 1) - 4) in a general way we introduce several notations and hypotheses.

1) Let  $H_M$ ,  $M-R, S$ , be two (real) Hilbert spaces with scalar products  $(\dots)_M$  (the induced norms are denoted by  $\|\cdot\|_M$ )

and let the Hilbert space  $H = H_R \times H_S$  (with element  $[v_R, v_S]$ ,  $v_R \in H_R$ ,  $v_S \in H_S$ ) have the scalar product  $(\cdot, \cdot)$  such that the norm  $\|v\| = (v, v)^{\frac{1}{2}}$  satisfies

$$C^{-1} |v| \leq \|v_R\|_R + \|v_S\|_S \leq C |v| \quad \forall v \in H$$

( $C$  here and in the sequel denotes a positive constant not necessarily the same at any two places). Further, let  $V \subset H$  be a separable reflexive Banach space normed by  $\|\cdot\|$ . Finally, the vector space  $V_M = \{\omega | \omega = v_M, v \in V\}$  ( $M=R, S$ ) and  $\overset{\circ}{V}_R = \{\omega | \omega = v_R, v \in V, v_S = 0\}$  should possess the following properties:  $V_M$  are subspaces of reflexive Banach spaces  $B_M \subset H_M$  normed by  $\|\cdot\|_M$ , it holds

$$C^{-1} \|v\| \leq \|v_R\|_R + \|v_S\|_S \leq C \|v\| \quad \forall v \in V,$$

$\bar{V}_R$ , the closure of  $V_R$  in  $B_R$ , is continuously imbedded in  $H_R$ , i.e.

$$|\omega_R| \leq C \|\omega\|_R \quad \forall \omega \in \bar{V}_R,$$

and  $\overset{\circ}{V}_R$  is dense in  $H_R$ .

Example. Let  $\Omega$ ,  $R$  and  $S$  be domains from section 1 with Lipschitz boundaries. We choose  $H_M = L^2(M)$ ,  $(u, v)_R = (\sigma u, v)_{L^2(R)}$  where  $\sigma \in L^\infty(R)$ ,  $\sigma \geq \sigma_0 > 0$ ,  $(u, v)_S = (u, v)_{L^2(S)}$ ,  $H = L^2(\Omega)$  ( $u_M$  is the restriction of  $u$  to  $M$ ),  $V = H_0^1(\Omega)$ ,  $\overset{\circ}{V}_R = H_0^1(R)$ ,  $\bar{V}_R = \{\omega | \omega \in H^1(R)$ ,  $\omega|_{\partial\Omega \cap \partial R} = 0\}$ ,  $B_R = H^1(R)$ ,  $\|\cdot\|_R = \|\cdot\|_{H^1(R)}$ ,  $\bar{V}_S = \{\omega | \omega \in H^1(S)$ ,  $\omega|_{\partial\Omega \cap \partial S} = 0\}$ ,  $B_S = H^1(S)$ ,  $\|\cdot\|_S = \|\cdot\|_{H^1(S)}$ .

Remark 5. We set  $H=H_R$  if  $H_S=\{0\}$ . The assumption 1) is to be understood as follows: There is a separable reflexive Banach space  $V$  normed by  $\|\cdot\|$  which is dense and continuously imbedded in  $H$ .

Remark 6. It is easy to see that  $\overset{\circ}{V}_R$  is a closed subspace of  $B_R$ . Further  $\overset{\circ}{V}_R$ ,  $\overline{V}_R$  and  $\overline{V}_S$ , being closed subspaces of reflexive Banach spaces  $B_R$  and  $B_S$ , respectively, are reflexive Banach spaces, and  $\overline{V}_R$  is dense in  $H_R$  because  $\overset{\circ}{V}_R \subset V_R$ .

We identify  $H_R$  with its dual by means of its scalar product  $(\dots)_R$ . Then  $H_R$  can be identified with subspaces of  $\overline{V}'_R$  and  $\overset{\circ}{V}'_R$  and we have inclusions

$$\overline{V}_R \subset H_R \subset \overline{V}'_R, \quad \overset{\circ}{V}_R \subset H_R \subset \overset{\circ}{V}'_R$$

where each space is dense in the following one and the injections are continuous. Furthermore, the scalar product  $\langle \dots \rangle_R$  in the duality between  $\overline{V}'_R$  and  $\overline{V}_R$  is an extension of  $(\dots)_R$ , i.e.

$$\langle u, v \rangle_R = (u, v)_R \quad \text{if } u \in H_R, v \in \overline{V}_R.$$

We denote the scalar product between  $V'$  and  $V$  by

$$\langle \dots \rangle$$

and between  $V'_S$  and  $V_S$  by

$$\langle \dots \rangle_S$$

Let  $A^M(u)$ ,  $m=R,S$ , be two, in general, nonlinear operators from  $\overline{V}_M$  to  $\overline{V}'_M$  with the following properties:

2)  $A^M(u)$  are hemicontinuous, i.e.  $\lambda \rightarrow \langle A^M(u+\lambda v), w \rangle_M$  are continuous functions on the interval  $(-\infty, \infty)$   $\forall u, v, w \in \overline{V}_M$ .

3) It holds

$$\|A^M(u)\|_* \leq C \|u\|_M^{p-1} \quad \forall u \in \bar{V}_M$$

where

$$1 < p < \infty.$$

4)  $A^M(u)$  are monotone, i.e.

$$\langle A^M(u) - A^M(v), u - v \rangle_M \geq 0 \quad \forall u, v \in \bar{V}_M$$

and  $A^S(u)$  is strictly monotone in the following sense:

$$\langle A^S(u) - A^S(v), u - v \rangle_S > 0 \quad \forall u, v \in \bar{V}_S, \quad u \neq v, \quad u - v \in \overset{\circ}{V}_S$$

where  $\overset{\circ}{V}_S = \{\omega | \omega = v_S, \quad v \in V, \quad v_R = 0\}$ .

The first of the above mentioned formulations is the following:

Problem P. Given

$$f^M \in L^{p'}(0, T; \bar{V}'_M), \quad M=R, S, \quad \text{and } u_0 \in H_R$$

find  $u \in W_R = \{u | u \in L^p(0, T; V); \quad u'_R \in L^{p'}(0, T; \bar{V}'_R)\}$  such that

$$(70) \quad \frac{du_R}{dt} + A^R(u_R) = f^R, \quad u(0)_R = u_0,$$

$$(71) \quad A^S(u_S) = f^S.$$

Remark 7. If  $H=H_R$  then we denote  $A^R(u)$  by  $A(u)$  and the assumptions 2, 3, 4 are to be understood as follows:  $A(u)$  is hemicontinuous, monotone and bounded, i.e.  $\|A(u)\|_* \leq C \|u\|^{p-1}$ . The formulation of the problem P reads: Given  $f \in L^{p'}(0, T; V')$  and  $u_0 \in H$  find  $u \in W = \{u | u \in L^p(0, T; V); \quad u' \in L^{p'}(0, T; V')\}$  such

that

$$\frac{du}{dt} + \Lambda(u) = f, \quad u(0) = u_0.$$

**Remark 8.** We could leave the requirement  $u'_R \in L^{p'}(0,T; \bar{V}'_R)$  because due to (70) it is automatically satisfied. From  $u \in W_R$  it follows  $u_R \in \{\omega | \omega \in L^p(0,T; \bar{V}_R); \omega' \in L^{p'}(0,T; \bar{V}'_R)\}$ . By lemma 2  $u_R \in C([0,T]; H_R)$  and the initial condition  $u(0)_R = u_0$  makes sense.

We introduce an equivalent variational formulation of problem P. To this end we define a form  $a(u,v)$  on  $V \times V$  which is linear in  $v$  and, in general, nonlinear in  $u$  and a functional  $f$  from  $L^{p'}(0,T; V')$ :

$$(72) \quad a(u,v) = \langle A^R(u_R), v_R \rangle_R + \langle A^S(u_S), v_S \rangle_S \quad \forall u, v \in V,$$

$$(73) \quad \langle f, v \rangle = \langle f^R, v_R \rangle_R + \langle f^S, v_S \rangle_S \quad \forall v \in V.$$

The form  $a(u,v)$  possesses the following properties:

a) it is hemicontinuous on  $V \times V$ , i.e.  $\lambda \rightarrow a(u + \lambda v, w)$  is a continuous function on the interval  $(-\infty, \infty)$   $\forall u, v, w \in V$ .

b)  $|a(u,v)| \leq C \|u\|^{p-1} \|v\| \quad \forall u, v \in V,$

c)  $a(u,v)$  is monotone on  $V \times V$ , i.e.

$$a(u, u-v) - a(v, u-v) \geq 0 \quad \forall u, v \in V.$$

At this place we add the last assumption which we shall later need:

$$5) \quad Q(v, v) \geq \alpha \|v\|^p \quad \text{or} \quad a(v, v) \geq \alpha [v]^p \quad \forall v \in V,$$

$$\alpha = \text{const} > 0.$$

Here  $[\cdot]$  is a seminorm on  $V$  such that

$$[v] + \lambda |v_R|_R \geq \beta \|v\| \quad \forall v \in V, \quad \lambda, \beta = \text{const} > 0.$$

**Problem P'.** Given  $f^M \in L^{p'}(0,T; \bar{V}_M')$ ,  $M=R,S$ , and  $u_0 \in H_R$  find  $u \in W_R$  such that

$$(74) \quad \frac{d}{dt} (u_R, \kappa_R)_R + a(u, \kappa) = \langle f, \kappa \rangle \quad \text{in } \mathcal{D}'((0,T)) \quad \forall \kappa \in V,$$

$$(75) \quad u(0)_R = u_0.$$

Here  $a(u,v)$  and  $f$  are defined by (72) and (73), respectively.

**Remark 9.** If  $H=H_R$  then the problem  $P'$  reads: Given  $f \in L^{p'}(0,T; V')$  and  $u_0 \in H$  find  $u \in W$  such that in  $\mathcal{D}'((0,T))$

$$\frac{d}{dt} (u, z) + a(u, z) = \langle f, z \rangle \quad \forall z \in V, \quad u(0) = u_0.$$

**Theorem 1.** Let the assumptions 1) and 3) be satisfied.  
Then the problems  $P$  and  $P'$  are equivalent.

**Proof.** If  $u$  is a solution of problem  $P$  then (70), (71), (72) and (73) imply

$$\left\langle \frac{du_R}{dt}, \kappa_R \right\rangle_R + a(u, \kappa) = \langle f, \kappa \rangle \quad \forall \kappa \in V.$$

All terms in this equation belong to  $L^{p'}(0,T)$  and for  $h(t) \in \mathcal{D}((0,T))$  we have  $\int_0^T \left\langle \frac{du_R}{dt}, \kappa_R \right\rangle_R h dt = - \int_0^T (u_R, \kappa_R)_R h' dt$  by lemma 2 as  $\kappa_R h' \in L^{p'}(0,T; \bar{V}_R')$ . Therefore, it holds (74).

Let  $u$  be a solution of problem  $P'$ . Choose  $\kappa = [\omega, 0]$ ,  $\omega \in V_R$  in (74). Then by (72) and (73)

$$\frac{d}{dt}(u_R, \omega)_R = \langle f^R - A^R(u_R), \omega \rangle_R \quad \text{in } \mathcal{D}'((0,T)) \quad \forall \omega \in V_R.$$

The function  $G(t) = (u(t)_R, \omega)_R$  is continuous on  $[0, T]$  because  $u_R \in C([0, T]; H_R)$  and the function  $g(t) = \langle f^R - A^R(u_R), \omega \rangle_R$  belongs to  $L^{p'}(0, T)$  (due to  $f^R, A^R(u_R) \in L^{p'}(0, T; \tilde{V}_R')$ ). Hence,

$F(t) = \int_0^t g(\tau) d\tau$  is an absolutely continuous function on  $[0, T]$ , consequently  $F = g$  a.e. and the distributional derivative of  $G - F$  is equal to zero (due to the above equation). Thus  $G(t) = C_0 + \int_0^t g(\tau) d\tau$  and evidently  $C_0 = G(0) = (u_0, \omega)_R$ . We have proved that

$$(u(t)_R, \omega)_R = (u_0, \omega)_R + \left\langle \int_0^t [f^R - A^R(u_R)], \omega \right\rangle_R dt \quad \forall \omega \in V_R.$$

As  $u(t)_R \in H_R \quad \forall t \in [0, T]$ ,  $u_0 \in H_R$  and  $V_R$  is dense and continuously imbedded in  $H_R$  it follows

$$u(t)_R = u_0 + \int_0^t [f^R - A^R(u_R)] d\tau \quad \text{taken as elements of } H_R.$$

Further,  $f^R - A^R(u_R) \in \tilde{V}_R'$  and  $H_R$  is dense and continuously imbedded in  $\tilde{V}_R'$ . Hence

$$u(t)_R = u_0 + \int_0^t [f^R - A^R(u_R)] d\tau \quad \text{taken as elements of } \tilde{V}_R'.$$

and by i) of lemma 1 it follows (70). Finally, as  $\frac{d}{dt}(u_R, \omega_R)_R = \langle u'_R, \omega_R \rangle_R$  the equations (74), (72) and (73) imply (71).

3. Now we define a completely discretized approximate solution of problem P. The discretization in space is carried out by means of a generalized Galerkin method (see Nečas [18],

p.47), in time we use the backward differentiation schemes mentioned in par.1. Written for the scalar equation  $\dot{y} = f$  these are

$$(76) \quad \sum_{j=0}^k \alpha_{k-j} y^{i-j} = \Delta t f^i$$

where

$$(77) \quad \alpha_0 = 1, \quad \alpha_0 = -1 \quad \text{if } k=1,$$

$$(78) \quad \alpha_2 = \frac{3}{2}, \quad \alpha_1 = -2, \quad \alpha_0 = \frac{1}{2} \quad \text{if } k=2.$$

We assume that there exists a family  $\{V^h\}$ ,  $h \in (0, h^*)$ ,  $h^* > 0$ , of finite dimensional subspaces of  $V$ , such that

$$(79) \quad \lim_{h \rightarrow 0^+} \text{dist}(V^h, v) = 0 \quad \forall v \in V.$$

We have three important remarks:

1) If a family  $\{V^{h_n}\}$ ,  $n=1, 2, \dots$ ,  $h_1 > h_2 > \dots$ ,  $\lim_{n \rightarrow \infty} h_n = 0$ , with  $\lim_{n \rightarrow \infty} \text{dist}(V^{h_n}, v) = 0 \quad \forall v \in V$  exists, then defining  $V^h = \bigcup_{n=1}^{\infty} V^{h_n}$  for  $h \in (h_{n+1}, h_n]$  we have a family with the above property.

2) A family  $V^h$  with the property (79) always exists under the assumption that  $V$  is a separable Banach space. In this case there exists a sequence  $\{\varphi_i\}_{i=1}^{\infty}$ ,  $\varphi_i \in V$ , such that for all  $n=1, 2, \dots$  the elements  $\varphi_1, \varphi_2, \dots, \varphi_n$  are linearly independent and the finite linear combinations of  $\varphi_i$ 's are dense in  $V$ . We take for  $V^{h_n}$ ,  $h_n = \frac{1}{m}$ , the space of all linear combinations of  $\varphi_1, \varphi_2, \dots, \varphi_m$ .

3) In case that  $V$  is a Hilbert space  $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ , and  $\Omega$  is a polyhedron, all in practice used finite element spaces have the property (79). We consider the boundary value problem: find  $z \in V$  such that  $a_0(z, \varphi) = a_0(v, \varphi) \quad \forall \varphi \in V$  where  $a_0(u, \varphi) = \int_{\Omega} \left[ \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} + u \varphi \right] dx$  and  $v$  is a given element

of  $V$  (of course,  $z=v$ ). If  $v_h$  is the finite element approximate solution and the finite element spaces satisfy certain requirements then  $\lim_{\Delta t \rightarrow 0} \|v - v_h\|_{H^1(\Omega)} = 0$  (see Ciarlet [19], Theorem 3.23, p.134);  $h$  is the maximum diameter of all elements.

We introduce  $\Delta t = \frac{T}{r}$ ,  $r$  being a natural number and consider the partition of the interval  $[0, T]$  with nodes

$$t_i = i \Delta t, \quad i=0, \dots, r.$$

We set

$$f^i = \frac{1}{\Delta t} \int_{t_{i-1}}^{t_i} f(\tau) d\tau \in V^*, \quad i = 1, \dots, r$$

and define  $U^i \in V^h$ ,  $i=1, \dots, r$  by

$$(80) \quad \left( \sum_{j=0}^k a_{k-j} U_R^{i-j}, \varphi_R \right)_R + \Delta t a(U^i, \varphi) = \Delta t \langle f^i, \varphi \rangle \quad \forall \varphi \in V^h,$$

$$U_R^0 = U_R^0 = u_0.$$

**Remark 10.** Instead of  $u_0$  we can take any approximation  $u_0^h$  of  $u_0$  such that  $\|u_0 - u_0^h\|_R \rightarrow 0$ .

In [17] it is proved that (80) is equivalent to a nonlinear system  $E(\underline{d})=0$ . Here  $E: \mathbb{R}^{d_h} \rightarrow \mathbb{R}^{d_h}$  (where  $d_h$  is the dimension of  $V^h$ ) is continuous, coercive and strictly monotone from which existence and uniqueness of  $U^i$  follows (see Ortega + Rheinbold [20], 6.4.2, 6.4.3). We extend the approximate solution  $U^\delta$ ,  $\delta = (\Delta t, h)$ , is the step function

$$(81) \quad U^\delta = U^i \text{ in } (t_{i-1}, t_i], \quad i=1, 2, \dots, r, \quad \delta = (\Delta t, h).$$

In [17] it is proved the following

**Theorem 2.** Let the assumptions 1) - 5) be fulfilled, let  $f^M \in L^p(0,T; V_M)$ ,  $M=R,S$ ,  $1 < p < \infty$ ,  $\frac{1}{p_0} + \frac{1}{p} = 1$ , and  $u_0 \in H_R$ . Then there exists a unique function  $u \in W_R = \{ u | u \in L^p(0,T; V); u'_R \in L^{p'}(0,T; \bar{V}_R') \}$  satisfying (74) and (75). Further, the approximate solution  $U^\delta$  defined by (80) and (81) exists, is unique and

$$(82) \quad U^\delta \rightarrow u \text{ in } L^p(0,T; V) \text{ weakly if } \delta \rightarrow 0.$$

If  $u \in C([0,T]; V)$  and the form  $a(u,v)$  is uniformly monotone, i.e.

$$a(u,u-v) - a(v,u-v) \geq \varphi(\|u-v\|) \quad \forall u, v \in V$$

where  $\varphi$  is a strictly increasing function on the interval  $[0, \infty)$  with  $\varphi(0)=0$ , then

$$(83) \quad \lim_{\delta \rightarrow 0} \|u_R - U_R^\delta\|_{C([0,T]; H_R)} = 0, \quad \lim_{\delta \rightarrow 0} \int_0^T \varphi(\|u_R - U_R^\delta\|) dt = 0.$$

**Remark 11.** If  $H=H_R$  then the assumptions 1) - 5) are the same as those of theorem 1.2 and 1.2 bis in Lions [21], p.162-163.

4. We apply theorem 2 to the problem (63)-(67). Let

$$\sigma \in L^\infty(R), \quad \sigma \geq \sigma_0 > 0$$

and let  $\partial\Omega, \partial R$  be polygons. We choose the spaces  $H_R, H_S$  etc. as in the example introduced at the beginning of paragraph 2. Then the assumption 1) is satisfied. We consider a regular family of triangulations  $\mathcal{T}_k$  (see Ciarlet [19], p.132) covering  $\Omega$  and satisfying the assumptions of theorem 3.2.3

from [19]. Then the family  $\{v^h\}$  satisfies the condition (79). The operators  $A^M(u_M)$  (in the sequel the subscript  $M=R,S$  means restriction to  $M$  and will be often left out) and the form  $a(u,v)$  are:

$$A^M(u_M) = - \sum_{i=1}^n \frac{\partial}{\partial x_i} (v_M \frac{\partial u_M}{\partial x_i}), \quad a(u,v) = \int_{\Omega} v \sum_{i=1}^n \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} dx.$$

Concerning the function  $v(x_1, x_2, \xi)$  we assume:

a)  $\forall \xi \in [0, \infty)$  the function  $(x_1, x_2) \rightarrow v(x_1, x_2, \xi)$

is measurable on  $\Omega$  and for almost all  $(x_1, x_2) \in \Omega$  the function  $\xi \rightarrow v(x_1, x_2, \xi)$  is continuous in  $[0, \infty)$  (Carathéodory's property);

b)  $\forall \xi \in [0, \infty)$  and for almost all  $(x_1, x_2) \in \Omega$   $v(x_1, x_2, \xi)$

is bounded from above and satisfies for almost all  $(x_1, x_2) \in \Omega$

$$(84) \quad \xi v(x_1, x_2, \xi) - \eta v(x_1, x_2, \eta) \geq \alpha(\xi - \eta) \quad \forall \xi \geq \eta \geq 0, \alpha = \text{const} > 0.$$

Then the assumptions 2)-4) are satisfied with  $p=2$  (see Gajewski+Grüger+Zacharias [1], p.68-71). (84) implies that

$$v(x_1, x_2, \xi) \geq \alpha > 0 \quad \text{for almost all } (x_1, x_2) \in \Omega \text{ and } \forall \xi \in [0, \infty).$$

Therefore the assumption 5) is also satisfied with  $p=2$  and, in addition,

$$a(u, u-v) - a(v, u-v) \geq \beta \|u-v\|_{H^1(\Omega)}^2, \quad \forall u, v \in H_0^1(\Omega), \beta > 0,$$

i.e.  $a(u, v)$  is uniformly monotone with  $g(\xi) = \beta \xi^2$ . Concerning the data  $J$  and  $u_0$  we require

$$J \in L^2(0, T; L^2(\Omega)), \quad u_0 \in L^2(\Omega).$$

The equation (80) can be written as follows:

$$(85) \quad \left\{ \begin{array}{l} \left( \sigma \sum_{j=0}^k \alpha_{k-j} U^{k-j}, \varphi \right)_{L^2(R)} + \Delta t \alpha(U^k, \varphi) - \Delta t(J^k, \varphi)_{L^2(\Omega)} \quad \forall \varphi \in V^k, \\ U_R^k = U_R^0 = u_0. \end{array} \right.$$

where  $J^k = \frac{1}{\Delta t} \int_{t_{k-1}}^{t_k} J(\cdot, t) dt.$

**Theorem 3.** Under the above introduced assumptions there exists a unique function  $u \in W_R$  which is the solution of the problem (63)-(67). Further, the approximate solution  $U^\delta$ , defined by (85) and (81) exists, is unique and

$$U^\delta \rightarrow u \text{ in } L^2(0, T; H_0^1(\Omega)) \text{ weakly if } \delta \rightarrow 0.$$

If  $u \in C([0, T]; H_0^1(\Omega))$  then

$$\lim_{\delta \rightarrow 0} \|u - U^\delta\|_{C([0, T], L^2(R))} = 0, \quad \lim_{\delta \rightarrow 0} \|u - U^\delta\|_{L^2(0, T; H_0^1(\Omega))} = 0.$$

Now we introduce error bounds under assumption that the solution  $u$  is enough smooth. We restrict ourselves to triangular elements and to piecewise linear trial functions which are mostly applied in practise. We take into account only triangulations which consist of triangles belonging either to  $\bar{R}$  or to  $\bar{S}$  and which form a regular family.

In applications, the coefficient  $V(x_1, x_2, \xi)$  is a piecewise continuous function of  $x = (x_1, x_2)$ . Every discontinuity in  $x$  along a boundary of a subdomain leads to a natural boundary condition of the form (66). We consider a model problem assu-

assuming  $v$  to be continuous in  $R$  and in  $S$  for all  $\xi \in [0, \infty)$  with discontinuity along  $\Gamma = \partial R \cap \partial S$ . We add two more assumptions:

$$|\xi v(x_1, x_2, \xi) - \eta v(x_1, x_2, \eta)| \leq L |\xi - \eta| \quad \forall \xi, \eta \in [0, \infty), \\ (x_1, x_2) \in R \cup S \\ J \in C([0, T]; L^2(\Omega)),$$

and investigate first the approximate solution constructed by means of the scheme (76), (77). The right-hand side of the defining equation will not be the same as in (85).  $U^i$  is now defined by

$$(86) \quad (\sigma \Delta U^i, \chi)_{L^2(R)} + \Delta t a(U^i, \chi) - \Delta t (J^i, \chi)_{L^2(\Omega)} \quad \forall \chi \in V^h$$

where  $U^i = U^{i-1} - U^{i-1}$  and  $J^i = J(., t_i)$ .

The initial condition is

$$(87) \quad U(0)_R = u_0^h$$

where  $u_0^h \in V_R^h = \{\omega | \omega = v_R, v \in V^h\}$  is any approximation of  $u_0$  such that

$$\|u_0 - u_0^h\|_{L^2(R)} \leq Ch \|u_0\|_{H^1(R)}.$$

Remark 12. If  $u_0 \in H^2(R)$  we can take for  $u_0^h$  the interpolate of  $u_0$ . If  $u$  satisfies (SS) then  $u^0$  must belong to  $H^1(R)$  and the orthogonal projection of  $u_0$  in  $L^2(R)$  onto the subspace  $V_R^h$  has the required property.

Theorem 4. Let the above assumptions be satisfied and let the exact solution  $u$  be so smooth that

$$(88) \quad u_M \in C([0,T]; H^2(M)), \quad M=R,S, \quad u \in L^2(0,T; H^1(\Omega)), \\ u_R'' \in L^2(0,T; \bar{V}_R').$$

Then for the approximate solution defined uniquely by (86) and (87) it holds

$$(89) \quad \left\{ \Delta t \sum_{i=1}^n \| u^i - U^i \|_{H^1(\Omega)}^2 \right\}^{1/2} = O(h + \Delta t).$$

Now we define  $U^i$  by means of the scheme (76), (78):

$$(90) \quad (\nabla [\frac{3}{2}U^i - 2U^{i-1} + \frac{1}{2}U^{i-2}], \varphi)_{L^2(R)} + \Delta t a(U^i, \varphi) = \Delta t (f^i, \varphi)_{L^2(\Omega)} \\ \forall \varphi \in V_R^h, \quad i \geq 2,$$

$$(91) \quad U_R^0 = \hat{u}_0^h, \quad U^1 \quad \text{computed from (86).}$$

Theorem 5. Let the exact solution fulfill (88) and

$$u_R'' \in C([0,T]; L^2(R)), \quad u_R''' \in L^2(0,T; \bar{V}_R').$$

Then for the approximate solution  $U^i$  defined uniquely by (90) and (91) it holds

$$(92) \quad \left\{ \Delta t \sum_{i=1}^n \| u^i - U^i \|_{H^1(\Omega)}^2 \right\}^{1/2} = O(h + \Delta t^2).$$

## VI. A damped nonlinear wave equation

Let  $\Omega \subset \mathbb{R}^N$  be a bounded domain with a boundary  $\partial\Omega$  and  $\{a_{ij}(x)\}_{i,j=1}^N$  be a uniformly positive definite matrix. Let  $d(x, t, u, z)$  and  $g(x, t, u, z)$  be piecewise continuous with respect to  $x$  and uniformly Lipschitz continuous with respect to  $t, u$  and  $z$  for  $(x, t) \in \bar{\Omega} \times [0, T]$  and  $u, z \in (-\infty, \infty)$ . Further, we assume

$$(93) \quad d(x, t, u, z) \geq 0.$$

We consider the equation

$$(94) \quad \frac{\partial^2 u}{\partial t^2} + d(x, t, u, u') \frac{\partial u}{\partial t} = Lu + g(x, t, u, u') \quad \text{in } \Omega$$

where

$$Lu = \sum_{i,j=1}^N \frac{\partial}{\partial x_i} [a_{ij}(x) \frac{\partial u}{\partial x_j}], \quad u' = \frac{\partial u}{\partial t},$$

with the boundary and initial conditions

$$(95) \quad u = 0 \text{ on } \partial\Omega \times (0, T), \quad u(x, 0) = u^0(x), \quad u'(x, 0) = \alpha^0(x) \quad \text{in } \Omega, \\ u^0, \alpha^0 \in H_0^1(\Omega).$$

We write the problem (94), (95) in a variational form.

We set

$$(96) \quad u' = z$$

so that  $z' = -d(x, t, u, z)z + Lu + g(x, t, u, z)$ . If the exact solution is smooth enough then it follows

$$(97) \quad (\kappa', v)_0 = -(d(x, t, u, z) \kappa, v)_0 - a(u, v) + (g(x, t, u, z), v)_0 \\ \forall v \in H_0^1(\Omega)$$

where

$$a(u, v) = \int_{\Omega} \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx.$$

The equations (96), (97) will serve as the starting point for the construction of the fully discrete approximate solution.

First, we define a semidiscrete solution. As in section III let  $\{v^h\}$ ,  $0 < h < h^*$ , be a family of finite dimensional subspaces of  $H_0^1(\Omega)$  possessing the approximation property (28).

By a semidiscrete Galerkin solution we mean a couple of functions  $U(x, t)$ ,  $Z(x, t) \in V^h$   $\forall t \in [0, T]$  satisfying in  $(0, T)$

$$(98) \quad U' \cdot Z, \quad (Z', v)_0 = -(d(x, t, U, Z)Z, v)_0 - a(U, v) + \\ + (g(x, t, U, Z), v)_0 \quad \forall v \in V^h, \quad U(x, 0) = U^0(x), \quad Z(x, 0) = Z^0(x).$$

Here  $U^0, Z^0 \in V^h$  are suitable approximations of  $u^0, z^0$ . The discretization in time is carried out by linear one or two-step A-stable methods defined by (15) and by the linearization procedure introduced in section III:

$$(99) \quad \sum_{j=0}^k \alpha_j U^{n+j} = \Delta t \sum_{j=0}^k \beta_j Z^{n+j}, \quad \left( \sum_{j=0}^k \alpha_j Z^{n+j}, v \right)_0 = -\Delta t \left( d \sum_{j=0}^k \bar{\beta}_j Z^{n+j}, v \right)_0 - \\ - \Delta t a \left( \sum_{j=0}^k \beta_j U^{n+j}, v \right) + \Delta t (g^n, v)_0 \quad \forall v \in V^h;$$

here

$$d^{\bar{n}} = d(x, t_{\bar{n}}, U^{\bar{n}}, Z^{\bar{n}}), \quad g^{\bar{n}} = g(x, t_{\bar{n}}, U^{\bar{n}}, Z^{\bar{n}})$$

and  $t_{\bar{n}}$ ,  $U^{\bar{n}}$ , and in the same way  $Z^{\bar{n}}$ , are given by (25), (26).

Let us show that the fully discrete approximate solution exists and is unique. Assuming that we have already computed  $U^k, Z^k, \dots, U^{n+k-1}, Z^{n+k-1}$  let us compute  $X = U^{n+k}, Y = Z^{n+k}$ . As  $U^i, Z^i, i=0, 1, \dots$  belong to  $V^h$  we may assume  $X$  and  $Y$  in the form  $X = \sum X_j v_j(x)$ ,  $Y = \sum Y_j v_j(x)$  where  $\{v_j(x)\}$  are basis functions of the finite dimensional space  $V^h$ . Denoting

$$\underline{X} = (X_1, X_2, \dots)^T, \quad \underline{Y} = (Y_1, Y_2, \dots)^T, \quad M = \{(v_i, v_j)_0\}_{i,j},$$

$$D^{\bar{n}} = \{(d^{\bar{n}} v_i, v_j)_0\}_{i,j}, \quad K = \{a(v_i, v_j)\}_{i,j}$$

we easily find out that  $\underline{X}$  and  $\underline{Y}$  are solutions of the following system of linear algebraic equations:

$$(100) \quad \alpha_k \underline{X} = \Delta t \beta_k \underline{Y} + \underline{a}, \quad \alpha_k M \underline{Y} = -\Delta t \beta_k D^{\bar{n}} \underline{Y} - \Delta t \beta_k K \underline{X} + \underline{b}.$$

Here  $\underline{a}, \underline{b}$  are known vectors, the matrices  $M$  and  $K$  are positive definite whereas  $D^{\bar{n}}$  is positive indefinite due to (93). From (100) we get

$$(101) \quad [\alpha_k M + \Delta t \beta_k (D^{\bar{n}} + \Delta t \alpha_k^{-1} \beta_k K)] \underline{Y} = \underline{c}, \quad \underline{X} = \Delta t \alpha_k^{-1} \beta_k \underline{Y} + \alpha_k^{-1} \underline{a}$$

( $\underline{c}$  is again a known vector). Evidently, the matrix  $\alpha_k M + \Delta t \beta_k (D^{\bar{n}} + \Delta t \alpha_k^{-1} \beta_k K)$  is positive definite which proves the above assertion.

The energy inequality (16) (used twice with  $b(u, v) = (u, v)_0$  as well as with  $b(u, v) = a(u, v)$ ) can be again successfully applied for deriving error estimates. We state the result for the case of  $\Theta$ -method with  $\Theta < \frac{1}{2}$  which is of order one ( $q=1$ ). Besides the hypotheses introduced above and besides some regularity conditions which we do not introduce we assume that  $U^0$  is the Ritz projection of  $u^0$ , i.e.  $a(U^0, v) = a(u^0, v) \quad \forall v \in V_h$  and that  $\|Z^0 - U^0\|_0 \leq Ch^{k+1}$  (e.g., we can take the interpolate of  $z^0$  in  $V_h$  for  $Z^0$ ). Then

$$(102) \quad \begin{aligned} \|u^m - U^m\|_0 &\leq C(h^{k+1} + \Delta t), \\ \|u^m - Z^m\|_0 &\leq C(h^{k+1} + \Delta t), \quad 1 \leq m \leq \Delta t^{-1}T, \\ \|u^m - U^m\|_1 &\leq C(h^k + \Delta t). \end{aligned}$$

## REFERENCES

- 1 J.D.Lambert, Computational Methods in Ordinary Differential Equations. Wiley, London, 1972.
- 2 G.A.Dahlquist, A Special Stability Problem for Linear Multistep Methods. BIT 3, 1963, pp.27-43.
- 3 W.Liniger, A Criterion for A-Stability of Linear Multistep Integration Formulae. Computing, Vol.3, 1968, pp.280-285.
- 4 M.Zlámal, Finite Element Methods in Heat Conduction Problems. The Mathematics of Finite Elements and Applications, vol.II (ed.J.R.Whiteman), Academic Press, London and New York, 1976, pp.85-104.
- 5 M.Zlámal, Finite Element Methods for Nonlinear Parabolic Equations. R.A.I.R.O., Anal.Num., vol.11, 1977, pp.93-107.
- 6 M.Crouzeix,P.-A.Raviart, Approximation des Problèmes d'Évolution. Lecture notes, 1980.
- 7 J.Douglas Jr. and T.Dupont, Galerkin Methods for Parabolic Equations. SIAM J.Numer.Anal., Vol.7, 1970, pp.575-626.
- 8 O.C.Zienkiewicz, The Finite Element Method. MacGraw-Hill, London, 1977.
- 9 M.F.Wheeler, A Priori  $L_2$ -Error Estimates for Galerkin Approximations to Parabolic Partial Differential Equations. SIAM J.Numer.Anal., Vol.10, 1973, pp.723-759.
- 10 T.Dupont, G.Fairweather and J.P.Johnson, Three-Level Galerkin Methods for Parabolic Equations. SIAM J.Numer.Anal., Vol.11, 1974, pp.392-410.
- 11 M.Lees, A Priori Estimates for the Solutions of Difference Approximations to Parabolic Differential Equations. Duke Math.J., Vol.27, 1960, pp.287-311.
- 12 A.Kufner, O.John, S.Fučík, Function Spaces, Academia, Prague, 1977.
- 13 R.Temam, Navier-Stokes Equations. North-Holland, 1977.
- 14 H.Gajewski, K.Gröger, K.Zacharias, Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Akademie-Verlag, Berlin, 1974.
- 15 V.Girault, P.-A.Raviart, Finite Element approximation of the Navier-Stokes Equations. Springer-Verlag, Berlin Heidelberg New York, 1979.

- 16 F.Melkes, M.Zlámal, Numerical Solution of Nonlinear Quasi-stationary Magnetic Fields. To appear.
- 17 M.Zlámal, Finite Element Solution of Nonlinear Quasistationary Magnetic Field. To appear.
- 18 J.Nečas, Les Méthodes Directes en Théorie des Équations Elliptiques. Academia, Prague, 1967.
- 19 P.G.Ciarlet, The Finite Element Method for Elliptic Problems. North-Holland, 1978.
- 20 J.M.Ortega, W.C.Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York and London, 1970.
- 21 J.L.Lions, Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires. Dunod, Gauthier-Villars, Paris, 1969.

DAT  
FILM